# 结合目标检测的人体行为识别

周 波 [1] 李俊峰 [1]

**摘 要** 人体行为识别领域的研究方法大多数是从原始视频帧中提取相关特征, 这些方法或多或少地引入了多余的背景信息, 从而给神经网络带来了较大的噪声. 为了解决背景信息干扰、视频帧存在的大量冗余信息、样本分类不均衡及个别类分类难的问题, 本文提出一种新的结合目标检测的人体行为识别的算法. 首先, 在人体行为识别的过程中增加目标检测机制, 使神经网络有侧重地学习人体的动作信息; 其次, 对视频进行分段随机采样, 建立跨越整个视频段的长时时域建模; 最后, 通过改进的神经网络损失函数再进行行为识别. 本文方法在常见的人体行为识别数据集 UCF101 和 HMDB51 上进行了大量的实验分析, 人体行为识别的准确率 (仅 RGB 图像) 分别可达 96.0% 和 75.3%, 明显高于当今主流人体行为识别算法.

**关键词** 深度学习, 行为识别, 卷积神经网络, 机器视觉, 目标检测

**引用格式** 周波, 李俊峰. 结合目标检测的人体行为识别. 自动化学报, 2020, **46**(9): 1961−1970

**DOI** 10.16383/j.aas.c180848

## Human Action Recognition Combined With Object Detection

ZHOU Bo[1] LI Jun-Feng[1]

**Abstract** Most of the research methods in the field of human action recognition extract relevant features from the original video frames. These methods introduce more or less redundant background information, which brings more noise to the neural network. In order to solve the problem of background information interference, large amount of redundant information in video frames, unbalanced sample classification and difficult classification of individual classes, this paper proposes a new algorithm for human action recognition combined with object detection. Firstly, the object detection mechanism is added in the process of human action recognition, so that the neural network has a focus on learning the motion information of the human body. Secondly, the video is segmentally and randomly sampled to establish long-term time domain modeling across the entire video segment. Finally, action recognition is performed through an improved neural network loss function. In this work, a large number of experimental analyses are performed on the popular human action recognition datasets UCF101 and HDBM51. The accuracy of human action recognition (RGB images only) is 96.0% and 75.3%, respectively, which is significantly higher than the state-of-the-art human action recognition algorithms.

**Key words** Deep learning, action recognition, convolutional neural network (CNN), computer vision, object detection

**Citation** Zhou Bo, Li Jun-Feng. Human action recognition combined with object detection. *Acta Automatica Sinica*, 2020, **46**(9): 1961−1970

目前, 人体行为分析成为一个十分活跃的计算机视觉领域, 包括对剪辑与未剪辑的视频段进行动作识别、时序动作提名、检测等研究方向分支. 人体行为识别在物联网与大数据的环境下具有广阔的应用场景, 包括体育运动、智能交通、虚拟现实、人机交互等领域. 由于人体行为的高复杂性与场景的多变化性[1], 使得行为识别成为一项非常具有挑战性的课题.

得益于卷积神经网络 (Convolutional neural network, CNN) 在图像处理领域取得的巨大成就以及大数据的发展, 目前基于深度学习的人体行为识别的方法[2−5] 已经优于基于经典的手工设计特征的方法[6−10], 且在三维空间的动作识别[11−14]领域也取得了显著成效.

然而, 基于深度学习的人体行为识别方法仍然存在一些难点[15]: 首先, Karpathy 等[16] 将单幅 RGB

收稿日期 2018-12-26 录用日期 2019-06-06
Manuscript received December 26, 2018; accepted June 6, 2019
国家自然科学基金 (61374022), 浙江省基础公益研究计划项目 (LGG18F030001), 金华市科学技术研究计划重点项目 (2018-1-027) 资助
Supported by National Basic Research Program of China (61374022), Zhejiang Basic Public Welfare Research Project (LGG18F030001), and Jinhua Science and Technology Research Program Key Project (2018-1-027)
本文责任编委 杨健
Recommended by Associate Editor YANG Jian
1. 浙江理工大学机械与自动控制学院自动化研究所 杭州 310018
1. Institute of Automation, Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018

图像作为深度学习模型的输入, 只考虑了视频的空间表观特征, 而忽视了视频与单幅静态图像的区别, 没有对视频的时域信息进行编码. 对此, Ji 等[17] 首次使用 3D-CNN 来获得运动信息; Donahue 等[18] 利用 2D-CNN 提取视频帧的表征信息, 紧接着连接一个长短期记忆 (Long short-term memory, LSTM) 循环神经网络或者 GRU (Gated recurrent unit) 等来学习帧与帧之间的运动信息[19]; 与 Donahue 等[18] 的做法不同, Zolfaghari 等[20] 将 2D-CNN 之后的循环神经网络替换成了 3D-CNN. Simonyan 等[21] 首次提出结合 RGB 图像与光流图像的双流卷积神经网络的方法, 利用视频相邻帧之间的信息差计算出光流作为网络的输入, 以期获得视频的时域信息. 后来的研究[22] 也表明: RGB 与光流的方法相融合可以提高在测试集上的精度. 对于 RGB + 光流的做法, 计算光流耗时也占用了计算机的额外内存. 所以, Tran 等[23] 提出一种基于 3D-CNN 的新的网络结构, 以期在单一网络中同时对视频的空域和时域信息进行编码, 而 3D-CNN 相比于 2D-CNN 的计算量较大.

其次, 不论是 2D-CNN 中堆叠的光流或是 3D-CNN 中堆叠的 RGB 图像, 都只对视频进行短期的时域信息编码, 尚未考虑视频的长时时域信息. 例如, 在一段视频中, 一个动作延续时间通常是几秒至几十秒甚至更长. 对此, Wang 等[24] 提出了时间段网络 (Temporal segment network, TSN), 一个输入视频被分为 $K$ 段 (segment), 而一个片段 (snippet) 从它对应的段中随机采样得到. 不同片段的类别得分采用段共识函数 (Segmental consensus function) 进行融合来产生段共识 (segmental consensus). 最后对所有模型的预测融合产生最终的预测结果.

另外, 针对视频中相邻两帧差异很小的情况, Zolfaghari 等[20] 提出 ECO (Efficient convolutional network for online video understanding) 以避免过多计算视频帧中的冗余信息, 从而实现实时动作识别. He 等[25] 为了提升模型在数据集上的准确度, 提出结合 RGB 图像、光流、音频信息的多模态融合方法, 此方法精度稍高但却十分占用计算空间与资源.

为了让 CNN 更好地学习到视频中的动作信息, 受目标检测算法的启发, 本文将区域候选网络 (Region proposal network, RPN) 应用于算法中, 将视频中人所在区域精确地提取出来, 变换到原图像大小, 以此作为神经网络的输入. 考虑到图像经过目标检测算法后得出的目标区域必定大小不一, 对此, 在本文算法中, 对每一幅图片做对齐操作, 确保输入到网络的图片大小一致. 此外, 类似于 TSN, 本文还对视频片段进行分段稀疏采样以使模型获得视频级的表达能力, 并将用于分类的交叉熵函数改进为 Lin 等[26] 提出的焦点损失 (Focal loss) 函数, 以解决分类问题中类别判断难以及可能存在的样本不均衡问题.

## 1 目标检测算法

综合目标检测算法的精度与速度, 本文采用 Ren 等[27] 提出的 Faster-RCNN 方法作为目标检测的框架. 首先, 每张图片经由特定特征提取网络提取特征, 得到的特征图经由区域候选网络生成约 $2k$ 个目标候选区域; 其次, $2k$ 个目标候选区域经过 ROI 池化层获得感兴趣的区域 (Region of interest, ROI), 感兴趣的区域经全连接层后产生两个分支, 经由 Bounddingbox regression 与 Softmax 输出分别得到目标所在原始图像区域的精准位置信息与其所属类别的概率; 最后, 对上述目标检测算法结果的两个信息进行调整, 得到对目标区域的裁剪图像与 warped 图像. Faster-RCNN 目标检测算法的具体流程如下.

### 1.1 特征提取

本文采用预训练的 VGG-Net 作为目标检测的特征提取网络提取视频帧的特征图, 其原理如图 1 所示. VGG-Net 有 13 个卷积层, 卷积核大小为 3×3, *padding* 值为 1, 卷积核水平与垂直移动步长为 1, 特征图大小与原始图像大小 ($W \times H$) 及卷积层参数关系为

$$W = \left\lfloor \frac{W - F + 2P}{S} + 1 \right\rfloor \tag{1}$$

$$H = \left\lfloor \frac{H - F + 2P}{S} + 1 \right\rfloor \tag{2}$$

每个卷积层后对应一个激活层, 激活层不改变图像大小, 所以原始图像经过卷积层和激活层后的特征图大小均不会改变. 4 个最大池化层对激活层输出进行 $2 \times 2$ 不重叠取最大值降采样, 所以输入的图像经过 VGG-Net 网络得到的特征图的长宽都为原始图像大小的 1/16. 最后得到的特征图为 512 维, 即特征图参数为 $(W/16) \times (H/16) \times 512$ 维度.

### 1.2 ROI 生成

本文采用的区域候选网络 (RPN) 如图 2 所示.

在 RPN 中, 输入的特征图经过 $kernelsize = 3 \times 3$, $padding = 1, stride = 1$ 卷积层与激活层, 大小维度仍不变, 再分别经过两个 1×1 的卷积层, 用于整合特征图不同维度的信息与降维. 位于上方的 1×1 卷积层输出 anchors 将用于二分类, 判断区域是否存在目标; 而位于下方的卷积层输出 anchors 用于做边框回归, 初步修正边框位置. 1×1 卷积后的特征图像素点映射到图片上的 3 种长宽比例和 3 种大小的区域, 以此生成 anchors. 最后, 由可能带有目标信息的 anchors 与初步修正的边框信息经过 ROI 池化层 (ROI pooling) 生成 ROI.

### 1.3　边框回归与类别预测

生成的 ROI 经过两个全连接层加激活层, 再分别进入两个不同的全连接层进行分类和边框回归, 输出 ROI 属于某一类的概率与精确的边框位置信息, 边框回归与类别预测流程如图 3 所示.

### 1.4　图像变换

通过目标检测算法可以得到目标的 Bounding

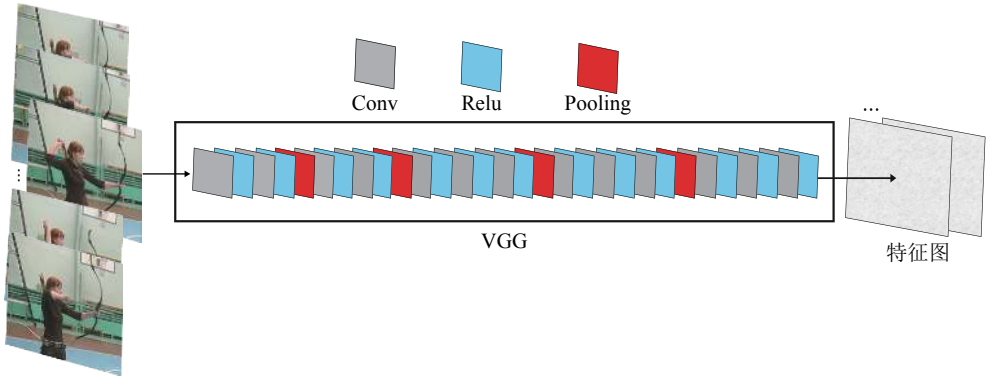



图 1　VGG 特征提取器

Fig. 1　VGG feature extractor

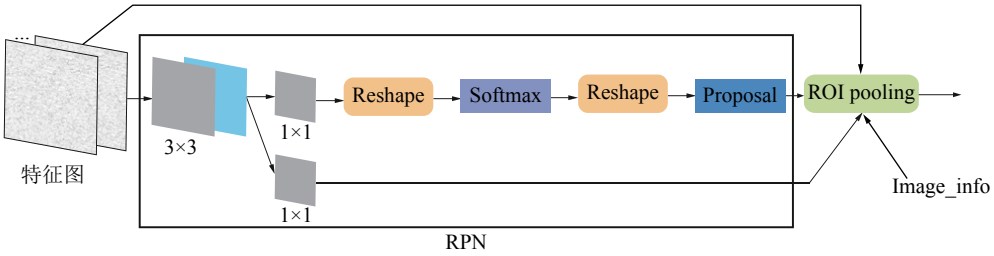



图 2　区域候选网络

Fig. 2　Region proposal network

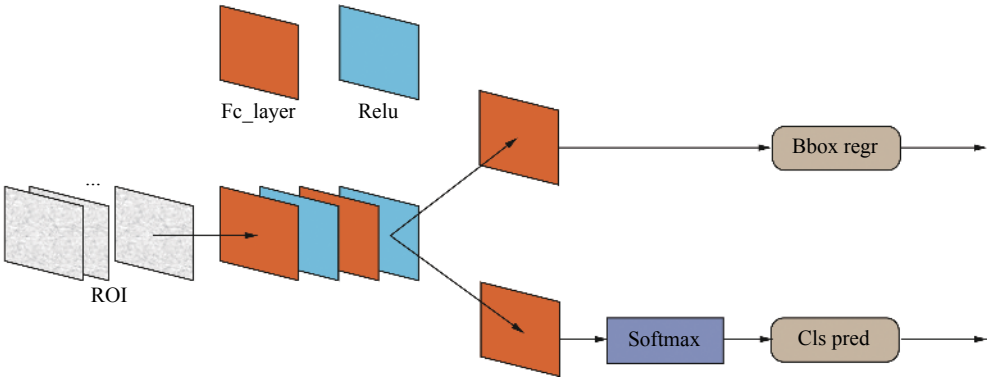



图 3　边框回归与类别预测

Fig. 3　Boundding box regression and class prediction

box 和对应的类别, 对非目标区域填充黑色得到 cropped 图像以及将目标区域扩充到原图大小得到 warped 图像. 而对未能找到图像中的人物信息或者所有目标 anchor 的总面积小于原图面积的 1/8 的图像, 取其原图作为训练样本. 采用此方法的原因是: 1) 理论上较小区域包含较少的图像信息; 2) 未检测到人物的图像可能会丢失主体部分信息. 考虑到人物目标在图像中大小不确定性与提取图像中的上下文必要背景信息, 本文将目标区域扩充至 coco 数据集中的 80 类, 形成以人物为主体的目标区域提取. 最后得到的结果如图 4 所示.

## 2 视频分段随机采样与训练

### 2.1 视频分段与采样

为了获得视频的长时时域信息, 建立视频级表达的 RGB 网络, 如图 5 所示, 在训练时对视频帧进行分段随机采样. 采用视频分段随机采样的原因是: 1) 堆叠的连续视频帧存在大量的冗余信息; 2) 许多方法都是基于局部推理的, 丧失获取持续时间长达数秒甚至数分钟的动作之间的相关关系. 本文提出方法类似于 TSN, 同样将视频帧分为 $K$ 段, 但与

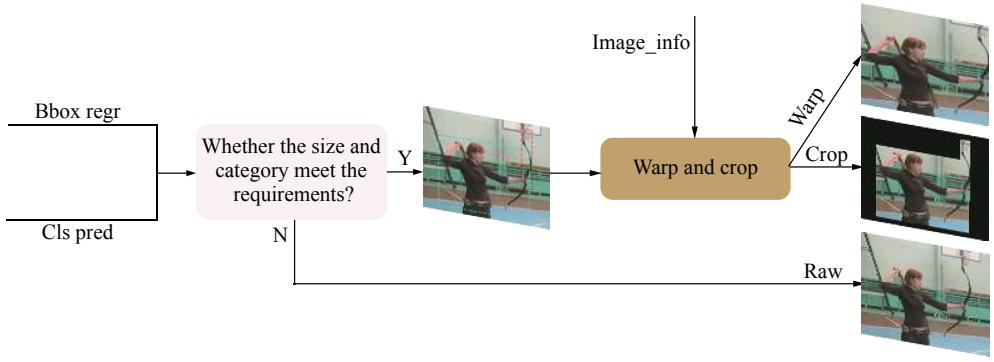



图 4 目标区域获取与图像变换

Fig. 4 Target area acquisition and image transformation

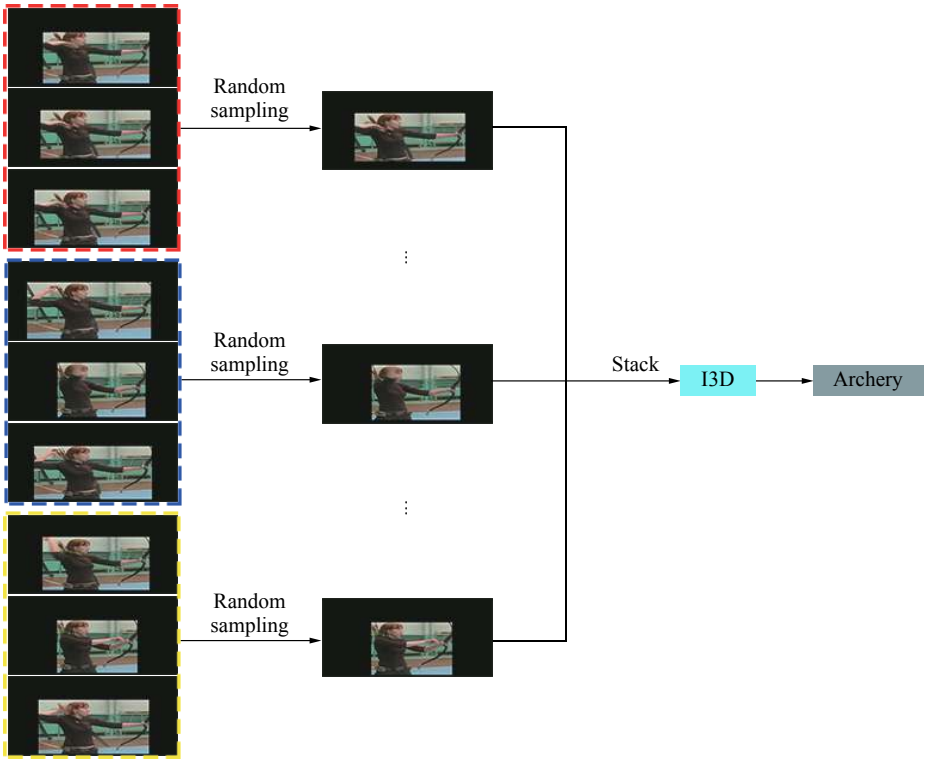



图 5 视频分段随机采样

Fig. 5 Video segmentation and random sampling

TSN 不同的是, TSN 将一个片段 (snippet) 从它对应的段中随机采样得到. 不同片段的类别得分采用段共识函数 (Segmental consensus function) 进行融合来产生段共识 (Segmental consensus), 然后对所有模式的预测融合产生最终的预测结果. 本文对每段采集 $N/K$ 帧图片, 将 $N$ 帧图片按时序顺序堆叠, 送入预训练的 I3D 网络中进行识别, 而并非每段视频对应一个模型, 然后进行模型融合.

### 2.2　I3D 网络结构

I3D 的实现, 将 Inception-v1 从 2D 扩展到 3D. 对于一个 2D 的模型, 将它的所有的 filters 和池化核增加一个时间维度, 例如将 $N \times N$ 的 filter 变成 $N \times N \times N$, 由 2D filters 得到 3D filters. 对 $N \times N$ 的 filter 重复复制 $N$ 遍, 再除以 $N$ 进行归一化. 确定感受野在空间、时间和网络深度的尺寸. 2D 网络与对应的 3D 网络在水平和竖直方向上的核大小和步长保持一致, 3D 网络在时间维度上的核大小和步长自由决定, 如果时间维度的感受野尺寸比空间维度的大, 将会合并不同物体的边缘信息. 反之, 将捕捉不到动态场景, I3D 网络结构如图 6 所示.

### 2.3　损失函数

Inception 框架中最后的损失函数为普通的交叉熵函数, $p$ 和 $y$ 分别为预测值与真实标签.

$$CE(p,y)=\begin{cases}-\ln(p), & \text{若 } y=1\\ -\ln(1-p), & \text{否则}\end{cases} \tag{3}$$

$$p_t=\begin{cases}p, & \text{若 } y=1\\ 1-p, & \text{否则}\end{cases} \tag{4}$$

且重写

$$CE(p,y)=CE(p_t)=-\ln(p_t)$$

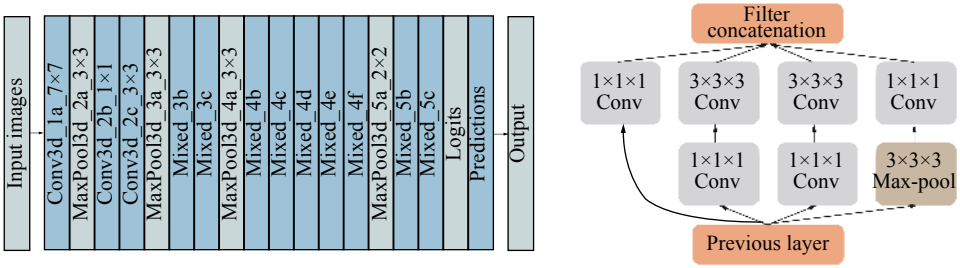
本文将其替换为 Focal loss 函数, 以处理样本分类难的问题, 转换后的损失函数为

$$FL(p_t)=-\alpha(1-p_t)^{\gamma}\ln(p_t) \tag{5}$$

Focal loss 函数可由式 (6) 和式 (7) 结合而成, 式 (6) 在交叉熵的基础上增加了调制参数 $\alpha$, $\alpha$ 的取值为: 当 $y=1$时, $\alpha=a$; 当 $y=-1$时, $\alpha=1-a$. 当正样本比例比负样本少很多时, 取 $a=0.5\sim1$ 来增大正样本对总的损失函数的权重. 这样即可解决正负样本不均衡问题.

$$CE(p_t)=-\alpha\ln(p_t) \tag{6}$$

从表 1 的实验结果来看 (本小节实验输入为: WI + RI, 加入了视频分段随机采样), Focal loss 函数的参数 $\alpha$ 对两个数据集的实验结果影响甚微. 但是, $\alpha=0.5$ 与 $\alpha=0.75$ 分别在 HMDB51 与 UCF101 数据集上较其他值有些微提升. 图 7 显示了 Focal loss 参数 $\alpha$ 的敏感曲线. 式 (7) 引入调制参数 $\gamma$, 当一个样本被分错的时候, $p_t$ 趋近于 0 时, $\gamma$ 趋近于 1, 与原不增加调制参数的损失相比, 损失




(a) I3D 网络结构
(a) Inflated inception network structure

(b) Inception 模块
(b) Inception modual

图 6　I3D 网络
Fig. 6　Inflated inception network

表 1　HMDB51 与 UCF101 数据集在不同 $\alpha$ 值下的实验结果 ($\gamma=1$) (%)
Table 1　Experimental results of HMDB51 and UCF101 data sets at different $\alpha$ values ($\gamma=1$) (%)

| HMDB51-FL- $\alpha$ | Split1 | Split2 | Split3 | Average | UCF101-FL- $\alpha$ | Split1 | Split2 | Split3 | Average |
|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 60.6 | 56.5 | 58.7 | 58.6 | 0.1 | 76.8 | 77.4 | 78.4 | 77.5 |
| 0.25 | 76.6 | 73.6 | 74.9 | 75.0 | 0.25 | 95.4 | 96.3 | 95.4 | 95.7 |
| 0.50 | **76.8** | 73.8 | **75.2** | **75.3** | 0.5 | 95.5 | 96.3 | **95.9** | 95.9 |
| 0.75 | 76.7 | **73.9** | 75.1 | 75.2 | 0.75 | **95.7** | **96.4** | 95.6 | 95.9 |
| 0.90 | 76.7 | 73.8 | 75.1 | 75.2 | 0.9 | 95.5 | 96.2 | 95.7 | 95.8 |
| 1.00 | 76.7 | 73.8 | 75.1 | 75.2 | 1 | 95.6 | 96.3 | 95.8 | 95.9 |

基本不变; 当 $p_t$ 趋近于 1 时, 此时样本分类正确且为易分类样本, $\gamma$ 趋近于 0, 意味着该类损失在总损失中权重很小.

$$CE(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{7}$$

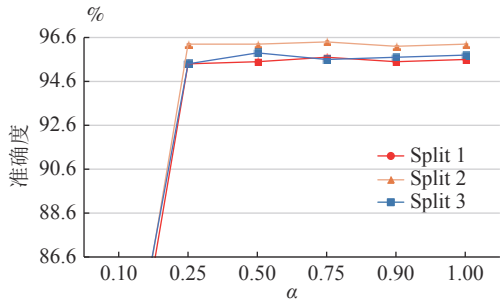
由于 HMDB51 与 UCF101 数据集实验精度对 Focal loss 函数的 $\alpha$ 参数不敏感, 故在本文中设置 $\gamma$ 由小到大进行实验. 表 2 显示了实验精度随 $\gamma$ 参数变化的规律. 图 8 显示了表 2 对应实验的直方图.

Focal loss 函数中的两个参数 $\alpha$ 和 $\gamma$ 相互协调进行控制.本文在 HMDB51 数据集上进行实验时采用的参数设置为 $\alpha = 0.5$, $\gamma = 5$; 在 UCF101 数据集上进行实验时采用的参数设置为 $\alpha = 0.75$, $\gamma = 5$.

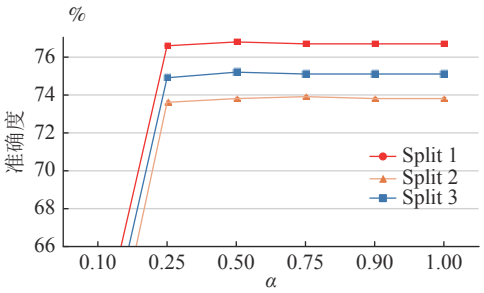
## 3 实验与分析

### 3.1 实验数据集

本文在最常见的行为识别数据集上评估所提出的网络架构, 主要包括比较受欢迎的数据集 UCF101




(a) UCF101 对参数 $\alpha$ 敏感曲线
(a) UCF101 sensitive curve to parameter $\alpha$




(b) HMDB51 对参数 $\alpha$ 敏感曲线
(b) HMDB51 sensitive curve to parameter $\alpha$

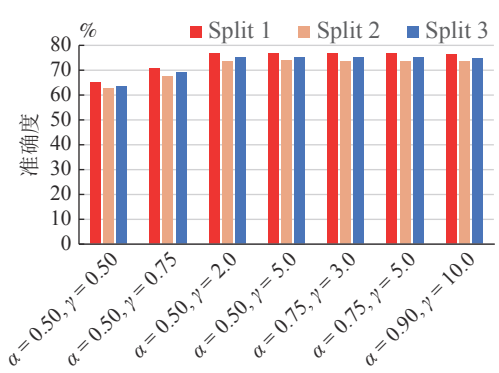
图 7 Focal loss 参数 $\alpha$ 敏感曲线

Fig. 7 Focal loss parameter $\alpha$ sensitivity curve

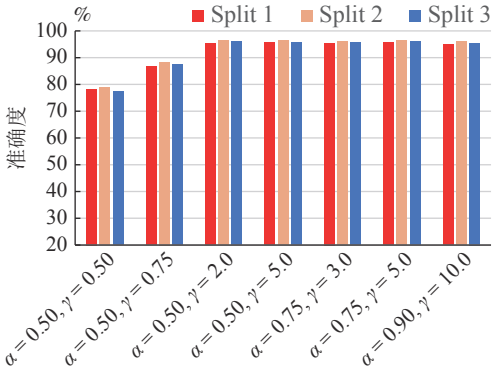
表 2 在 Focal loss 的不同参数值条件下的实验精度对比 (%)

Table 2 Comparison of experimental precision under different parameter values of focal loss (%)

| HMDB51 | Split 1 | Split 2 | Split 3 | Average | UCF101 | Split 1 | Split 2 | Split 3 | Average |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha=0.50,\ \gamma=0.50$ | 65.3 | 62.8 | 63.5 | 63.9 | $\alpha=0.50,\ \gamma=0.50$ | 78.3 | 78.9 | 77.4 | 78.2 |
| $\alpha=0.50,\ \gamma=0.75$ | 70.8 | 67.5 | 69.2 | 69.2 | $\alpha=0.50,\ \gamma=0.75$ | 86.8 | 88.4 | 87.4 | 87.5 |
| $\alpha=0.50,\ \gamma=2.00$ | 76.6 | 73.7 | 75.1 | 75.1 | $\alpha=0.50,\ \gamma=2.00$ | 95.4 | 96.3 | **96** | 95.9 |
| $\alpha=0.50,\ \gamma=5.00$ | **76.9** | **73.8** | **75.3** | **75.3** | $\alpha=0.50,\ \gamma=5.00$ | 95.6 | 96.3 | 95.8 | 95.9 |
| $\alpha=0.75,\ \gamma=3.00$ | 76.7 | 73.7 | 75.2 | 75.2 | $\alpha=0.75,\ \gamma=3.00$ | 95.5 | 96.2 | 95.7 | 95.8 |
| $\alpha=0.75,\ \gamma=5.00$ | 76.7 | 73.7 | 75.1 | 75.2 | $\alpha=0.75,\ \gamma=5.00$ | **95.7** | **96.4** | 95.9 | **96** |
| $\alpha=0.90,\ \gamma=10.0$ | 76.3 | 73.4 | 74.7 | 74.8 | $\alpha=0.90,\ \gamma=10.0$ | 95 | 95.9 | 95.5 | 95.5 |




(a) HMDB51 在不同 Focal loss 参数条件下的实验精度直方图
(a) Experimental precision histogram of HMDB51 under different focal loss parameters




(b) UCF101 在不同 Focal loss 参数条件下的实验精度直方图
(b) Experimental precision histogram of UCF101 under different focal loss parameters

图 8 不同 Focal loss 参数条件下实验精度直方图

Fig. 8 Experimental precision histogram under different focal loss parameters

和 HMDB51, 以便将其性能与目前主流的方法进行比较.

UCF101 数据集是从 YouTube 收集的具有 101 个动作类别的逼真动作视频的动作识别数据集, 此数据集是 UCF50 数据集的扩展. 凭借来自 101 个动作类别的 13 320 个视频, UCF101 在动作方面提供了最大的多样性, 并且存在相机运动、物体外观和姿势、物体比例、视点、杂乱背景、照明条件等较大的变化, 它是迄今为止依然具有一定挑战性的数据集. 101 个动作类别中的视频分为 25 组, 每组可包含 4 ~ 7 个动作视频. 来自同一组的视频可能共享一些共同的功能, 例如类似的背景、类似的观点等. 动作类别可以分为 5 种类型: 1) 人–物体相互作用; 2) 仅身体动作; 3) 人–人相互作用; 4) 演奏乐器; 5) 运动.

HMDB51 数据集内容主要来自电影, 一小部分来自公共数据库, 如 Prelinger 存档、YouTube 和 Google 视频. 该数据集包含 6 849 个剪辑, 分为 51 个动作类别, 每个动画类别至少包含 101 个剪辑. 操作类别可以分为 5 种类型: 1) 一般的面部动作微笑; 2) 对象操纵的面部动作; 3) 一般身体动作; 4) 与对象互动的身体动作; 5) 人体互动的身体动作.

### 3.2　实验条件

实验计算机配置为 Intel Core i5-8500@ 3.0 GHz, NVIDA GeForce 1080 TI GPU, 操作系统为 Windows 10. 实验中, 卷积神经网络基于 Tensorflow 平台设计实现. 网络训练采用小批量随机梯度下降法, 动量为 0.9, 权值在每 10 个 epoch 衰减 1 次, 衰减率为 0.1, HMDB51 数据集的批大小为 6, UCF101 数据集的批大小为 8. 采用在 ImageNet + Kinetics 行为库上预训练的 Inception 3D 网络, 初始学习率设为 0.001.

### 3.3　实验结果与分析

表 3 显示了本文算法在行为识别数据集 UCF101 和 HMDB51 上使用不同输入图像、Warped 图像与 Cropped 图像的识别结果. 分别对数据集划分的 3 个子数据集进行训练, 测试准确度, 最后对所有测试集结果取平均.

实验结果表明, Warped 图像比 Cropped 图像具有更高的可辩别性, 原因在于 Cropped 图像比 Warped 图像多了黑色区域. 而事实上每个类的图片因 anchor 大小及比例不固定的原因都存在这样的黑色区域. 因此, 相同分辨率的 WI 图像比 CI 图像具有更少的冗余信息和更多的有效信息. WI + RI 图像相对原始图片而言, 在减少噪声的同时, 扩大了人体动作区域在图像中的所占比重, 使得训练结果有所提升.

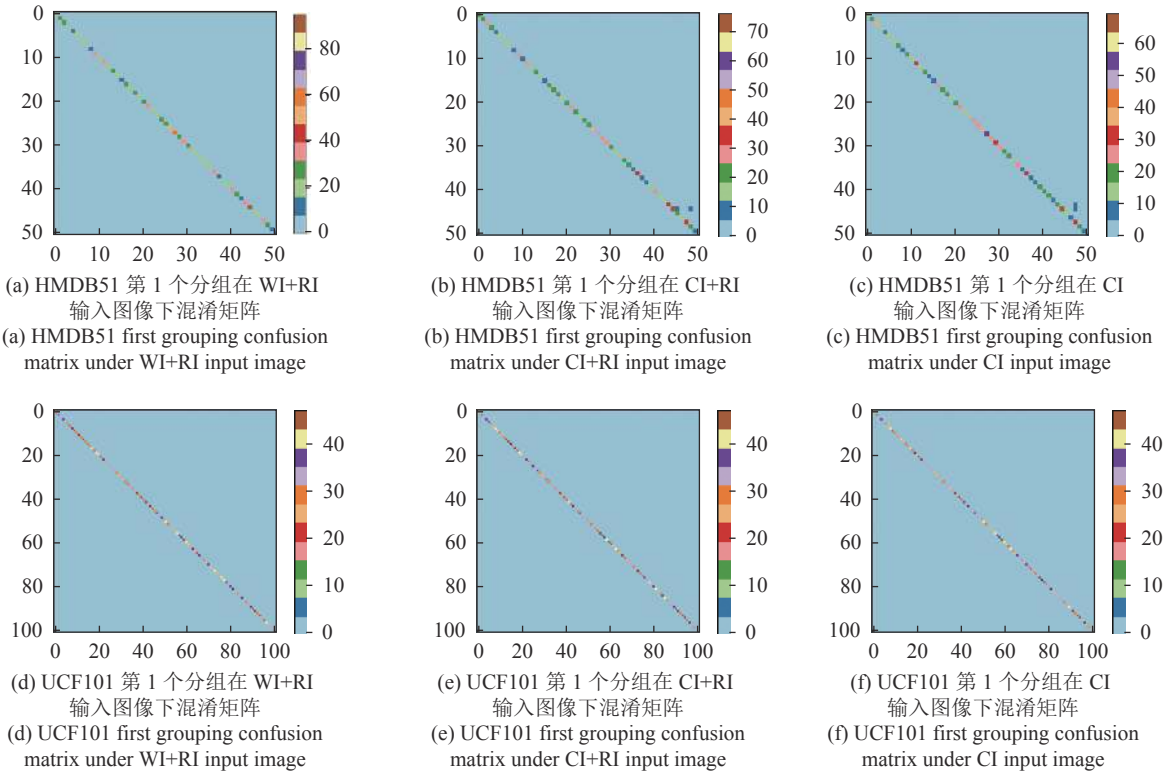
图 9 显示了 UCF101 和 HMDB51 数据集的不同类别图像在第 1 个分组的测试集上的混淆矩阵图, UCF101 数据集因预测准确率较高无法直观地从混淆矩阵中看出模型预测各类别时准确率的差异; 而 HMDB51 数据集可以明显地看出, 在第 48 类, 49 类 WI + RI 的预测概率分别比后两者有显著提升.

图 10 显示了不同输入图像下的 I3D 网络一些类别的测试精度对比. 包含两个数据集上的 WI + RI 较 CI + RI 与 CI 预测概率提升最大的类别、最平稳的的类别以及下降最大的类别. HMDB51 第 48 类 throw 位于提升最大类别之中, 与图 7 的混淆矩阵相符. 两个数据集上相对提升最大的类别是, eat, throw, fall_floor, kayaking, bowling, frisbeecatch. 这些行为相对右边的行为而言背景占据较大范围且与行为相关性强. 行为相对下降最多的是 shoot_ball, laugh, shake_hands, lunges, shavingbeard, mixing. 这些行为相对人体占据图像小或动作幅度不大, 所以完全去除背景能够更有效提升该行为的识别率.

表 4 显示了本文提出的算法与现有其他算法在行为识别数据集 UCF101 和 HMDB51 上的对比结果. 在不对输入进行分段随机采样且不采用 Focal loss 函数的情况下, 实验结果显示利用目标检测算法能够有效地学习视频中人物的动作信息并加以辨别. 本文用 WI + RI 的图像输入形式在删减过多

表 3　UCF101 与 HMDB51 数据集实验结果 (%)
Table 3　Experimental results of UCF101 and HMDB51 (%)

| UCF101-Input | Split 1 | Split 2 | Split 3 | Average | HMDB51-Input | Split 1 | Split 2 | Split 3 | Average |
|---|---|---|---|---|---|---|---|---|---|
| CI | 87.6 | 91.7 | 90.9 | 90.1 | CI | 71.3 | 67.1 | 68.8 | 69.7 |
| WI | 90.4 | 92.2 | 92.5 | 91.7 | WI | 74.1 | 70.2 | 70.6 | 71.6 |
| RI | 95.2 | 95.8 | 95.4 | 95.5 | RI | 75.9 | 73.1 | 75.0 | 74.7 |
| CI+RI | 91.7 | 92.7 | 92.9 | 92.4 | CI+RI | 73.3 | 71.8 | 72.0 | 72.4 |
| WI+RI | 95.7 | 96.4 | 96.0 | **96.0** | WI+RI | 76.8 | 73.9 | 75.3 | **75.3** |



(a) HMDB51 第 1 个分组在 WI+RI 输入图像下混淆矩阵
(a) HMDB51 first grouping confusion matrix under WI+RI input image

(b) HMDB51 第 1 个分组在 CI+RI 输入图像下混淆矩阵
(b) HMDB51 first grouping confusion matrix under CI+RI input image

(c) HMDB51 第 1 个分组在 CI 输入图像下混淆矩阵
(c) HMDB51 first grouping confusion matrix under CI input image

(d) UCF101 第 1 个分组在 WI+RI 输入图像下混淆矩阵
(d) UCF101 first grouping confusion matrix under WI+RI input image

(e) UCF101 第 1 个分组在 CI+RI 输入图像下混淆矩阵
(e) UCF101 first grouping confusion matrix under CI+RI input image

(f) UCF101 第 1 个分组在 CI 输入图像下混淆矩阵
(f) UCF101 first grouping confusion matrix under CI input image

图 9 混淆矩阵

Fig. 9 Confusion matrix

表 4 不同算法在 UCF101 和 HMDB51 数据集上识别准确率对比 (%)

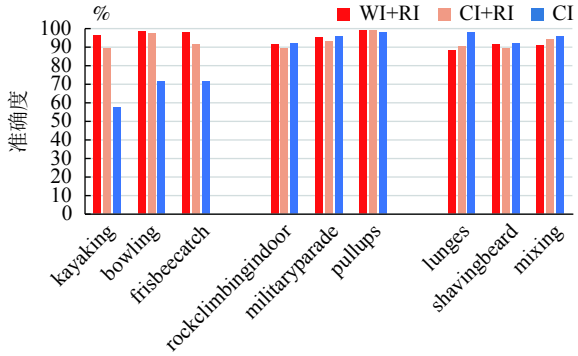
Table 4 Comparison with the state-of-the-art on UCF101 and HMDB51 (%)

| 算法 | Pre-training | UCF101 | HMDB51 |
|---|---|---|---|
| LTC[28] | Sports-1M | 82.4 | 48.7 |
| C3D[23] | Sports-1M | 85.8 | 54.9 |
| TSN[24] | ImageNet | 86.4 | 53.7 |
| DTPP[29] | ImageNet | 89.7 | 61.1 |
| C3D[5] | Kinetics | 89.8 | 62.1 |
| T3D[30] | Kinetics | 91.7 | 61.1 |
| ARTNet[31] | Kinetics | 94.3 | 70.9 |
| TSN[24] | ImageNet+Kinetics | 91.1 | – |
| I3D[2] | ImageNet+Kinetics | 95.6 | 74.8 |
| PM without TS & FL | ImageNet+Kinetics | 95.8 | 95.1 |
| PM without FL | ImageNet+Kinetics | 95.9 | 75.1 |
| PM without TS | ImageNet+Kinetics | 95.9 | 75.2 |
| Proposed method (all) | ImageNet+Kinetics | **96.0** | **75.3** |

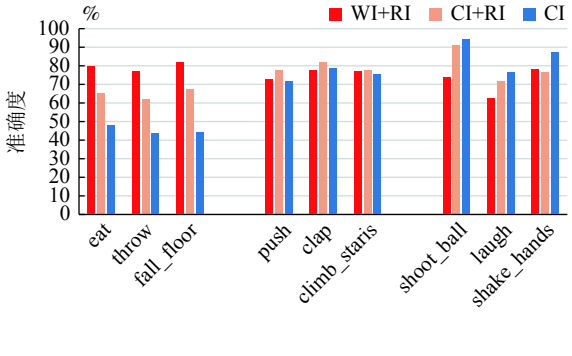
背景信息与保留必要的背景信息中取得平衡, 有效地提高了行为识别的准确率. 而消融实验则表明, Focal loss 函数与视频分段随机采样策略进一步提高了本文算法的竞争力.

## 4 结论

本文提出了一种结合目标检测的人体行为识别方法. 通过在人体行为识别算法中加入目标检测机制, 使神经网络能够有侧重地学习人体的动作信息, 而减弱部分不必要的背景噪声干扰, 同时对不合要求的图像进行替换, 达到平衡背景取舍的作用. 结合视频分段随机采样, 改进 I3D 网络的损失函数. 本文提出的算法在常用数据集上进行实验, 并与其他先进算法进行比较, 体现出了良好的性能, 实验




(a) UCF101 在不同的输入图像下部分类别精度对比
(a) UCF101 partial category accuracy comparison under different input images




(b) HMDB51 在不同的输入图像下部分类别精度对比
(b) HMDB51 partial category accuracy comparison under different input images

图 10　不同的输入图像下 I3D 网络测试精度对比

Fig. 10　Comparison of I3D network test accuracy under different inputs

结果验证了本文提出方法的有效性.

## References


1 Zhu Hong-Lei, Zhu Chang-Sheng, Xu Zhi-Gang. Research advances on human activity recognition datasets. *Acta Automatica Sinica*, 2018, **44**(6): 978−1004
(朱红蕾, 朱昶胜, 徐志刚. 人体行为识别数据集研究进展. 自动化学报, 2018, **44**(6): 978−1004)

2 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 4724−4733

3 Ng Y H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: Deep networks for video classification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 4694−4702

4 Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and imagenet? In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 6546−6555

5 Tran D, Ray J, Shou Z, Chang S F, Paluri M. Convnet architecture search for spatiotemporal feature learning. arXiv: 1708.05038, 2017.

6 Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV). Sydney, Australia: IEEE, 2013. 3551−3558

7 Dalal N. Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886−893

8 Chaudhry R. Ravichandran A. Hager G. Vidal R. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA: IEEE, 2009. 1932−1939

9 Knopp J, Prasad M, Willems G, Timofte R, VanGool L. Hough transformand 3D SURF for robust threedimensional classification. In: Proceedings of the 11th European Conference on Computer Vision (ECCV2010). Berlin Heidelberg, Germany: Springer. 2010. 589−602

10 Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 2013, **105**(3): 222−245

11 Yang Y H, Deng C, Gao S L, Liu W, Tao D P, Gao X B. Discriminative multi-instance multi-task learning for 3d action recognition. *IEEE Transactions on Multimedia*, 2017, **19**(3): 519−529

12 Yang Y H, Deng C, Tao D P, Zhang S T, Liu W, Gao X B. Latent max-margin multi-task learning with skelets for 3d action recognition. *IEEE Transactions on Cybernetics*, 2017, **47**(2): 439−448

13 Kim T S, Reiter A. Interpretable 3d human action analysis with temporal convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA: IEEE, 2017. 1623−1631

14 Yang Y, Liu R S, Deng C, Gao X B. Multi-task human action recognition via exploring super-category. *Signal Process*, 2016, **124**: 36−44

15 Zhu Yu, Zhao Jiang-Kun, Wang Yi-Ning, Zheng Bing-Bing. A review of human action recognition based on deep learning. *Acta Automatica Sinica*, 2016, **42**(6): 848−857
(朱煜, 赵江坤, 王逸宁, 郑兵兵. 基于深度学习的人体行为识别算法综述. 自动化学报, 2016, **42**(6): 848−857)

16 Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F F. Large-scale video classification with convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014. 1725−1732

17 Ji S W, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(1): 221−231

18 Donahue J, Hendricks L A, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **39**(4):

677−691

19 Cho K, Van Merrienboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv: 1409.1259, 2014.

20 Zolfaghari M, Singh K, Brox T. ECO: Efficient convolutional network for online video understanding. arXiv: 1804.09066, 2018.

21 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. *Advance in Neural Information Processing Systems*, 2014, **1**(4): 568−576

22 Sevilla-Lara L, Liao Y Y, Guney F, Jampani V, Geiger A, Black M J. On the integration of optical flow and action recognition. arXiv: 1712.08416, 2017.

23 Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4489−4497.

24 Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O, Van Gool L. Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, the Netherlands: Springer, 2016. 20−36

25 He D L, Li F, Zhao Q J, Long X, Fu Y, Wen S L. Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. arXiv: 1806.10319, 2018.

26 Lin T Y, Goyal P, Girshick R, He K M, Dollár P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2999−3007

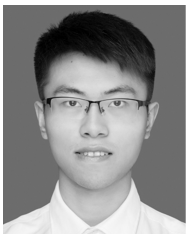
27 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **39**(6): 1137−1149

28 Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **40**(6): 1510−1517

29 Zhu J G, Zou W, Zhu Z. End-to-end video-level representation learning for action recognition. In: Proceedings of the 24th International Conference on Pattern Recognition (ICPR). Beijing, China, 2018. 645−650

30 Diba A, Fayyaz M, Sharma V, Karami A H, Arzani M M, Yousefzadeh R, et al. Temporal 3d convnets: New architecture and transfer learning for video classification. arXiv: 1711.08200, 2017.

31 Wang L M, Li W, Li W, Van Gool L. Appearance-and-relation networks for video classification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 1430−1439



**周 波** 浙江理工大学硕士研究生. 2017 年获浙江理工大学机械与自动控制学院学士学位. 主要研究方向为深度学习, 计算机视觉与模式识别. E-mail: zhoubodewy@163.com

(**ZHOU Bo** Master student at the Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University. He received his bachelor degree from Zhejiang Sci-Tech University in 2017. His research interest covers deep learning, computer vision, and pattern recognition.)

**李俊峰** 浙江理工大学机械与自动控制学院副教授. 2010 年获得东华大学工学博士学位. 主要研究方向为图像质量评价, 人体行为识别, 产品视觉检测. 本文通信作者.
E-mail: ljf2003@zstu.edu.cn

(**LI Jun-Feng** Associate professor at the Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University. He received his Ph.D. degree from Donghua University in 2010. His research interest covers image quality assessment, human action recognition, and product visual inspection. Corresponding author of this paper.)