结合分水岭和回归网络的视频时序动作选举算法

黄韵文[1]，王斐[2*]，李景宏[1]，王国锐[2]

[1]东北大学信息科学与工程学院，辽宁 沈阳 110819；

[2]东北大学机器人科学与工程学院，辽宁 沈阳 110169

摘要 针对时序动作选举任务，设计一种两段式动作候选区域选举网络。第一段将改进的分水岭算法应用于一维时序信号，通过浸水聚类产生多种不同长度的候选区域，实现动作时序边界的粗定位，进而提出一种时序金字塔结构化方法，引入动作片段的上下文信息模块，对候选区域的主体信息和上下文信息进行结构化建模，生成一个增强的全局特征。第二段利用时序坐标回归算法定位动作边界，同时加入动作/背景分类器过滤背景候选区域，得到更加精确的时序边界。整个网络以三维卷积神经网络(C3D)提取的单元级特征进行训练，挖掘了视频时域和空域的丰富语义，在提升算法精度的同时大大提升了训练效率。在两大基准数据集 Thumos 14 和 ActivityNet 上进行测试，结果表明，与已有方法相比，两段式视频时序动作选举算法达到了最优平均召回率，可有效提高动作定位的精度。

关键词 机器视觉；视频时序检测；动作定位；金字塔池化；时序上下文

中图分类号 TP391 **文献标识码** A **doi**：10.3788/CJL201946.1109001

Algorithm for Video Temporal Action Proposal Combining Watershed and Regression Networks

Huang Yunwen[1], Wang Fei[2*], Li Jinghong[1], Wang Guorui[2]

[1]*College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning* 110819, *China*;

[2]*Faculty of Robot Science and Engineering, Northeastern University, Shenyang, Liaoning* 110169, *China*

Abstract A two-stage action-candidate regional proposal network is designed herein for a temporal action detection task. The first stage applies a modified watershed algorithm to an one-dimensional temporal signal to form candidate regions with different lengths by immersion clustering, which obtains a rough localization of action temporal boundary. Then, a temporal pyramid structural method is introduced to model the structure of action instances and their contextual information, generating an enhanced global feature. The second stage performs a temporal-coordinate regression algorithm to local the action boundary, and simultaneously a classifier for the action and boundary is added to filter the candidate regions of background for obtaining a more accurate temporal boundary. Furthermore, an unit-level feature extracted by a three-dimensional convolution neural network (C3D) is used to train the entire two-stage proposal algorithm, which contains both spatial and temporal information and considerably improves training efficiency while improving the accuracy of the algorithm. Experiments on two large-scale benchmark datasets, Thumos 14 and ActivityNet, show that the proposed approach achieves the optimal average recall rate over other state-of-the-art methods, indicating that this method can efficiently improve the precision of an action localization task.

Key words machine vision; video temporal detection; action localization; pyramid pooling; temporal context

OCIS codes 150.1135; 100.4996; 100.3008; 150.0155

1 引 言

视频中的时序动作检测任务是当今计算机视觉领域最具挑战性的研究课题之一，具有视频监控与安全、人机交互、视频检索和医疗监控等应用[1-2]。与视频动作识别相比，时序动作检测任务更具挑战性，不仅需要预测动作类别，还要预测动作发生时精确的开始和结束时间点[2-5]。

收稿日期：2019-06-04；**修回日期**：2019-06-20；**录用日期**：2019-07-05

* **E-mail**：wangfei@mail.neu.edu.cn

随着卷积神经网络的发展，视频动作识别的准确度得到显著提升[3]。传统的针对动作片段的时序检测方法[4]缺乏对动作上下文信息的利用，导致时序检测的精度并不理想，并且大多依赖手工特征，Thumos Challenge 2014[1]和 2015[5]中最佳的方法都使用改进的密集轨迹(iDT)和 Fisher 矢量。近年来，一些研究将 iDT 特征与帧级深度网络[1]提取的外观特征相结合，然而这样的 2D 卷积网络只能捕获空间表征信息，不能捕获视频中对于建模动作来说非常重要的运动信息。

2D 图像的目标检测任务[6-7]与视频时序动作检测任务密切相关，传统的目标检测方法大多依赖于密集滑动窗口[8]。近年提出的空间结构建模[9]是目标检测任务的关键步骤，RoI 池化的引入可以以最小的额外成本对目标的空间结构建模。

在时序检测领域，基于视频片段的方法[3]通常是独立处理各片段而不考虑它们之间的时序结构。现有方法在两个关键方面受到限制：1)SCNN[10]等方法只能对固定且有限长度的时序结构建模，没有顾及到动作发生长度的多样性；2)可以对较长动作建模的循环网络[11]等方法依赖于密集的时序片段采样，计算量太大，无法实现端到端训练。

针对上述问题，本文主要进行以下工作：1)提出一个两段式候选区域选举模型。第一段选举算法首次将改进的分水岭算法应用于一维时序信号，通过聚类产生多种不同长度的候选区域，兼顾动作长度的多样性，实现动作边界的粗定位；第二段用时序回归选举网络设计两个同级结构，该结构能够同时实现动作边界的坐标回归定位及动作/背景候选区域的分类和过滤，得到精确的时序边界，实现端到端的训练。2)提出时序金字塔结构化分析方法，引入动作片段的上下文信息模块，对候选区域的主体信息和上下文信息进行结构化建模，生成一个全局表达，针对性地增强了动作片段的时序结构特征，允许对不同长度动作片段的建模。3)整个框架基于三维卷积神经网络(C3D)[4]单元级特征进行训练。C3D 能够同时挖掘视频的时域和空域信息，提升时序检测的精度。相比于传统的单帧训练，单元级特征训练在大大提升时序检测速度的同时保证了精度。提出的时序动作选举算法生成的候选区域在 Thumos 14 和 ActivityNet 中达到了最优的平均召回率，在时序定位应用中具有极具竞争力的平均精度均值。

2　改进的分水岭选举算法

结合分水岭和回归网络的视频时序动作选举算法整体结构如图 1 所示。整个网络流程如下：1)将视频以单元级为单位(n_u=16 frame)输入到视频特征提取器 C3D 中，提取单元级特征 $\boldsymbol{f}_u$；2)将视频帧以单元级为单位输入到改进的分水岭算法中浸水聚类，得到初级候选区域 p；3)将初级候选区域 p 用 $\boldsymbol{f}_u$ 表示，结合上下文信息模块将 p 扩展增强为 p^*，p^* 经过时序金字塔结构化处理，生成全局区域特征 $\boldsymbol{D}_c$；4)将 $\boldsymbol{D}_c$ 作为回归/分类网络的输入，训练得到候选区域精确的时序区间 $\boldsymbol{P}_m$ 及候选区域的动作/背景分类。

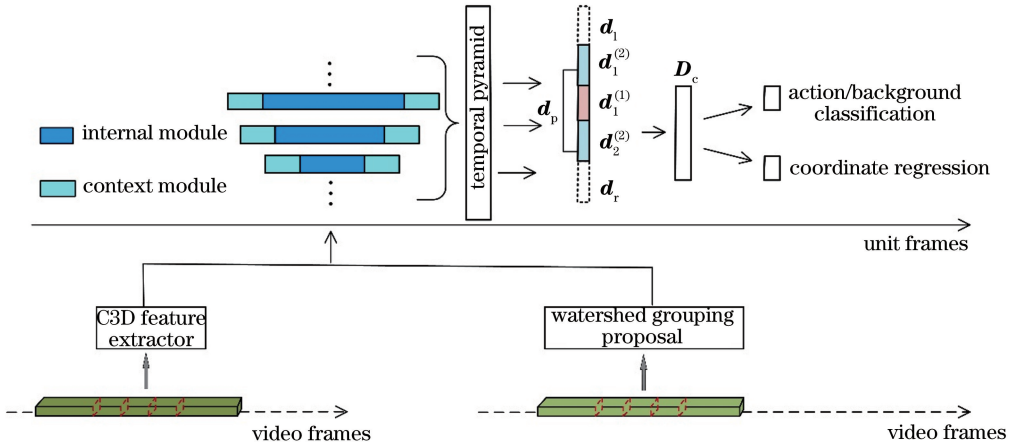

图 1　时序检测整体结构图

Fig. 1　Whole architecture of temporal detection algorithm

2.1　C3D 单元级特征处理

计算高效性对于大型视频分析任务非常重要。近期的研究[10-11]将双流网络或三维卷积网络应用于视频特征提取，虽然效率有所提升，但仍是针对单帧图片的特征提取，效率无法得到进一步提高。针对此问题，提出一种基于单元级特征训练的方法，不再

以单帧为单位提取视频特征。一个包含 T 帧图片的视频 $V=\{t_i\}_1^T$ 被切分为 T/n_u 个连续视频单元，其中 t_i 为第 i 帧图片，i 为图片的序号，n_u 为一个视频单元包含的视频帧数。将一个视频单元表示为 $u=\{t_i\}_{s_f}^{n_u}$，其中 s_f 是视频单元起始帧，s_f+n_u 是视频单元结束帧。将视频单元输入视频特征提取器 $\boldsymbol{E}_V$ 中，得到单元级特征 $\boldsymbol{f}_u=\boldsymbol{E}_V(u)$。为同时挖掘视频的空域和时域特征，对单元帧间的运动信息建模，选用 C3D 作为视频特征提取器。采用文献[4]的网络结构，将 C3D 的 fc6 输出作为单元级特征，fc6 为 C3D 的第 6 层，即全连接层的输出。

2.2　改进的分水岭选举算法

将应用于二维图像分割任务的分水岭算法改进为应用于一维时序信号聚类的时序区域选举算法。在二维图像分割领域，分水岭算法的优点是计算简单，对于较均匀的连通目标有较好的分割效果，而视频中动作发生的时间段相对于整个视频时间段来说，也是较均匀的联通目标。同时，受文献[12]空间动作定位任务中二元动作性分类器的启发，本文创将二元动作性分类器应用于一维时序动作定位任务，主要思想是筛选出动作可能性得分高的单元动作片段，将包含高分单元动作片段较多的时序区间汇聚为一个连通的候选区域，作为对动作时序边界的粗定位。

将二元动作性分类器[12]应用于时序动作定位任务，二元动作性分类器采用在 Kinetics 数据集下预训练的 ResNet。利用 ResNet 对每个单元级片段进行动作可能性评估，将其包含动作真值的概率作为该片段的动作可能性得分。

改进的分水岭算法将视频以单元级片段（$n_u=16$ frame）为基础单位输入二元动作性分类器中，产生连续的动作可能性信号（如图 2 上部曲线），将信号取负值映射为一维时序补充信号（图 2 下部曲线），作为分水岭算法的输入。动作可能性评分高或低的时序片段被分别映射为“盆地”或“山峰”。在此，地貌以不同注水等级 γ 浸水，得到一组盆地 $G(\gamma)$。

基于改进的分水岭方法得到盆地 $G(\gamma)$。依次从各个种子盆地出发浸水，当盆地时序区间长度 L_{basin} 与该浸水区间总长度 $L_{overall}$ 的比值小于阈值 τ 时，即 $\dfrac{\sum L_{basin}}{L_{overall}}<\tau$ 时，停止浸水。如图 2 所示，吸收的盆地和山峰被聚类为连通的候选区域，每个虚线框代表一个动作候选区域。

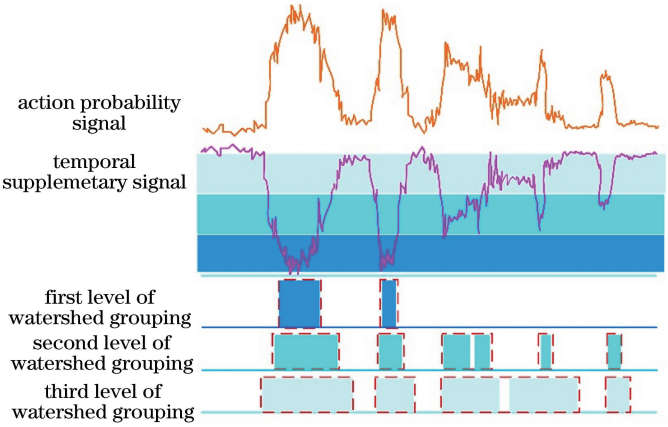

图 2　改进的分水岭选举算法原理图

Fig. 2　Principle of improved watershed proposal algorithm

考虑到视频动作时序长度的多样性，选取多组 τ 和 γ 生成候选区域并集，τ 和 γ 均为(0,1)之间以 0.05为间隔的采样值。最后，对并集应用非极大值抑制算法，过滤掉重叠率高于 0.95 的候选区域，剩余的候选区域即为动作粗定位候选区域。

区别于传统的滑窗法[10]大批量随机选取候选区域，改进的分水岭选举算法依据动作可能性对时序区域进行更合理的初步定位，过滤掉一部分背景信息，节省了后续的计算成本。不同注水等级 γ 的聚类可得到多种长度的动作候选区域，克服了传统算法可建模动作长度单一的问题。

3　时序回归选举算法

3.1　时序金字塔结构化算法

将视频单元输入到视频特征提取器 C3D，可得到单元级特征 $\boldsymbol{f}_u$。而一个动作候选区域包含多个视频单元，即包含多组 $\boldsymbol{f}_u$，将一个候选区域表示为 $p=\{u_j\}_{b_u}^{e_u}$，其中，u 为视频单元，j 为第 j 个视频单

元，b_u 为起始视频单元，e_u 为结束视频单元，$e_u=b_u+n_p$，n_p 为候选区域 p 包含视频单元的个数。

为更好地利用视频中的运动信息和因果语义信息，引入上下文信息模块，$p_l=\{u_j\}_{b_u-n_{ctx}}^{b_u}$ 和 $p_r=\{u_j\}_{e_u}^{e_u+n_{ctx}}$ 分别代表动作候选区域 p 之前和之后的上下文信息模块，n_{ctx} 代表上下文信息模块包含的视频单元个数，此时，一个动作候选区域 $p=\{u_j\}_{b_u}^{e_u}$ 被扩展为 $p^*=\{u_j\}_{b_u-n_{ctx}}^{b_u+n_{ctx}}$。将原候选区域 $p=\{u_j\}_{b_u}^{e_u}$ 称为内部候选区域，将上下文信息模块 p_l 和 p_r 称为扩展候选区域。

时序金字塔结构化算法的设计灵感来自于 spacial pyramid pooling[9] 在目标识别任务中的成功应用。对于一个扩展增强过的候选区域 p^*，将 p^* 划分为内部候选区域 p 和扩展候选区域 p_l、p_r。首先分别对这 3 个区域进行时序金字塔池化，提取区域跨度特征 $\boldsymbol{d}_p$，$\boldsymbol{d}_l$，$\boldsymbol{d}_r$，再将这 3 个区域跨度的特征融合到一起，得到一个全局区域金字塔特征 $\boldsymbol{D}_c$，计算公式为

$$\boldsymbol{D}_c=\boldsymbol{D}_{l=1}(p_l)\,\|\,\boldsymbol{D}_{l=2}(p)\,\|\,\boldsymbol{D}_{l=1}(p_r),\tag{1}$$

该公式可以更详细地表示为

$$\boldsymbol{D}_c=\boldsymbol{D}_{l=1}\{\{u_j\}_{b_u-n_{ctx}}^{b_u}p_l\}\,\|\,\boldsymbol{D}_{l=2}\{\{u_j\}_{b_u}^{e_u}\}\,\|\,\times\boldsymbol{D}_{l=1}\{p_r=\{u_j\}_{e_u}^{e_u+n_{ctx}}\},\tag{2}$$

式中，$\|$ 表示向量拼接，$\boldsymbol{D}_{l=1}$ 和 $\boldsymbol{D}_{l=2}$ 分别代表 $l_{level}=1$ 和 $l_{level}=2$ 的金字塔池化操作。对于一个 L-level 的时序金字塔，在第 l 个 l_{level}，区域跨度特征被均匀地划分为 B_l 个部分。将第 l 个 l_{level} 的第 j 个部分表示为 $[b_{lj},e_{lj}]$，则它的池化特征提取公式为

$$\boldsymbol{d}_j^{(l)}=\frac{1}{|e_{lj}-b_{lj}+1|}\sum_{t=b_{lj}}^{e_{lj}}\boldsymbol{f}_u。\tag{3}$$

串联各部分 $\boldsymbol{d}_j^{(l)}$，得到区域跨度特征 $\boldsymbol{D}_l=(\boldsymbol{d}_j^{(l)}\,|\,l=1,\cdots,L;j=1,\cdots,B_l)$，其中，$L$ 为池化 level 的总等级数，B_l 为区域跨度特征被均匀划分后的第 l 个部分。由于内部候选区域 p 反映动作本身，包含更丰富的动作信息，因此，对 p 进行更精细的 $l_{level}=2$ 的金字塔池化（$L=2$，$B_1=1$，$B_2=2$）；而对扩展候选区域 p_l、p_r 作 $l_{level}=1$ 的金字塔池化（$L=1$，$B_1=1$）。对于一个扩展增强的动作候选区域 p^*，金字塔结构化算法如图 3 所示。

3.2　单元级时序回归选举网络

在时序金字塔结构化模块之后连接两个同级网络结构，一个为动作/背景分类器 C，另一个为时序坐标回归器 R。分类器 C 与回归器 R 组成一个支持端到端多任务训练的架构，称为时序回归选举网络。网络以单元级特征为基础单位进行坐标回归，保证了网络训练的效率。单元级时序回归选举网络有两个主要优点：1）以单元级（$n_u=16$ frame）为基础单位进行坐标回归，单元级特征网络比帧级特征网络更易于训练和学习；2）没有采用参数化坐标，而是直接学习两个坐标变量的修正值——起始坐标修正值和结束坐标修正值。

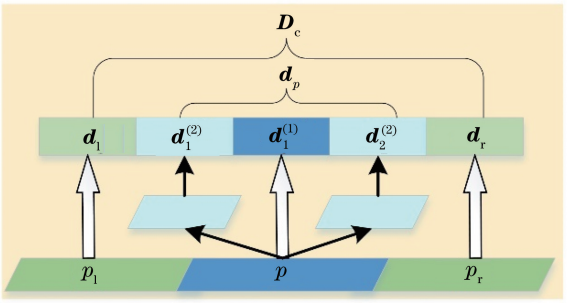

图 3　内部区域及扩展区域的时序金字塔结构化

Fig. 3　Structure of temporal pyramid of internal and extended regions

将全局区域金字塔特征 $\boldsymbol{C}_c$ 作为两个同级网络的输入，动作/背景分类器 C 的输出层是一个 softmax 层，输出此候选区域包含动作片段的置信分数，实现对背景候选区域的过滤。对于一个增强的候选区域 p^*，时序坐标回归器 R 以与其最相近的真值 gt 的时序区间为目标，回归器的输出是起始时序坐标和结束时序坐标的修正值，可表示为

$$R_b=b_{proposal}-b_{gt},R_e=e_{proposal}-e_{gt},\tag{4}$$

式中，$b_{proposal}$ 和 $e_{proposal}$ 分别代表扩展增强的候选区域 p^* 的起始单元级特征和终止单元级特征，b_{gt} 和 e_{gt} 分别代表与之匹配的真值的起始单元级特征和终止单元级特征。

训练阶段收集两种训练样本。1）动作样本：将与最相近的真值 gt 的时间交并比（tIoU）大于 0.5 的候选区域标记为正样本；2）背景样本：将与真值没有任何重叠，即 tIoU 为 0 的候选区域标记为负样本。采用多任务损失函数实现分类器 C 和回归器 R 的同时训练，即

$$L=L_{cls}+\lambda L_{reg},\tag{5}$$

式中，L_{cls} 是 softmax 分类损失函数，L_{reg} 是时序坐标回归器 R 的损失函数，λ 是可调节超参数。L_{reg} 的计算公式为

$$L_{reg}=\frac{1}{N_{pos}}\sum_{i=1}^{N}[l_i^*=p_{pos}]\times\left|(R_{b,i}-R_{b,i}^*)+(R_{e,i}-R_{e,i}^*)\right|,\tag{6}$$

式中：l_i^* 是候选区域的标签（动作/背景）；N 是动作

候选区域的总数；p_{pos}是指该候选区域是动作区域；[·]是指示函数，当候选区域为正样本时，$[l_i^*=\mathrm{pos}]=1$，否则等于0；N_{pos}是正样本的数量；$R_{\mathrm{b},i}$和$R_{\mathrm{e},i}$是时序坐标回归器预测的回归修正值；$R_{\mathrm{b},i}^*$和$R_{\mathrm{e},i}^*$是时序区间真正应该修正的回归修正真值。

将改进的分水岭选举算法与时序回归选举算法相组合，分别实现对动作时序区间的粗定位和细定位，然后加入上下文信息模块、C3D单元级特征提取模块和时序金字塔结构化模块，得到一个完整的两段式时序区域选举网络，整个架构支持端到端的训练。

4　实验与分析

4.1　数据集与实验设置

Thumos 14数据集的时序动作定位子集包含来自20个类别的超过20 h的视频。此子集包含200个验证集视频和213个测试集视频。由于Thumos 14的训练集仅包含修剪的视频，本文模型在验证集上训练。

ActivityNet数据集提供了丰富的视频数据，该数据集共包含3个子集，分别是v1.1、v1.2、v1.3。其中v1.3被用作ActivityNet Challenge的竞赛数据集，包含200个类别共19994个视频。每个子集都以2∶1∶1的比例被划分为训练集、验证集和测试集。

参数设置。1）C3D单元级特征：以$n_{\mathrm{u}}=16$ frame为间隔，将视频图片输入到Sports1 m预训练的C3D模型中，采用文献[4]的网络结构，将fc6层输出作为视频的单元级特征$\boldsymbol{f}_{\mathrm{u}}$。2）改进的分水岭选举算法：二元动作性分类器ResNet以文献[13]的方法在Kinetics数据集下预训练，再以$n_{\mathrm{u}}=16$ frame为间隔将视频帧输入到预训练好的ResNet，得到单元级的动作性得分，作为改进的分水岭选举算法的输入信号，经过分水岭算法聚类，得到动作发生时间不同长度的初步定位结果集合，对集合采用非极大值抑制算法，过滤掉重叠率高于0.95的候选区域，得到初级候选区域；3）时序回归选举网络训练：一个小批次样本数为128，上下文特征单元数$n_{\mathrm{ctx}}=4$，$\lambda=2$，中间全连接层维数$f_{\mathrm{m}}=1000$，选用Adam优化器以学习率$\alpha=0.005$进行训练。

4.2　两段式候选区域选举网络

为验证两段式时序区域选举算法“改进的分水岭选举算法和时序回归选举算法”生成的候选区域的质量，设计了两实验。

1）从候选区域是否精准的角度，将本文算法与其他时序动作选举方法的召回率进行比较，采用2种评估方法。a）AR-AN：随着平均检索到的候选区域数量增加，平均召回率相应地变化；b）Recall-AN-tIoU：固定选取前平均数（AN，A_{N}）检索到的候选区域，随着tIoU的变化，召回率相应的变化。

2）从对后续时序定位效果影响的角度，将不同选举方法产生的候选区域连接到相同的动作分类器上。根据最终定位坐标的平均精度均值（mAP），评价其产生的候选区域的质量。其中相同的分类器选用两种，一种是SVM分类器，一种是SCNN中的第三段结合时序重叠率的分类器。a）SVM分类器：利用C3D的fc6特征训练，实现21个类别的分类（20个动作类，1个背景类）；b）SCNN定位器：SCNN[10]三段式时序动作检测方法中的第三阶段，是结合时序重叠率的新型分类器，由分类器的动作预测得分对候选区域进行非极大值抑制过滤，产生最终定位坐标。

4.2.1　两段式区域选举算法候选区域的精度

采用AR-AN和Recall-AN-tIoU方法下对两段式区域选举算法产生的候选区域与其他时序区域选举算法（如DAPs[14]，Sparse-prop[15]，SST[16]，BSN[17]）产生的候选区域进行比较，结果在AR-AN和Recall-AN-tIoU方法下比较，如图4所示。

从图4(a)可以看出，当选定候选区域数量相同时，两段式区域选举算法产生的候选区域的平均召回率平稳，且高于DAPs、Sparse-prop、SST、BSN。其中：SST的思想类似于本文改进的分水岭选举算法，同样是将改进的二维目标检测方法应用于一维时序空间，但召回率远低于两段式区域选举算法；BSN同样注重处理时序动作的上下文信息，引入了起始边界特征和结束边界特征，但精确度略低于两段式区域选举算法。从图4(b)可以看出，在tIoU值较高的情况下，两段式区域选举算法仍可保持较高的召回率。这说明两段式区域选举算法产生的候选区域普遍更精确。

4.2.2　两段式候选区域对时序定位的影响

将本文算法与其他先进的时序动作选举算法，包括DAPs[14]、Sparse-prop[15]、SST[16]、SCNN-prop[10]、BSN[17]，在Thumos 14数据集上进行对比。将不同选举算法产生的候选区域输入到相同的动作分类器中，通过最终定位的mAP取0.5来衡量候选区域的质量。同时，为验证“分水岭＋时序回归”两个选举阶段各自的有效性，对改进的分水岭选举算法（Watershed）、时序回归选举算法（Reg）、整

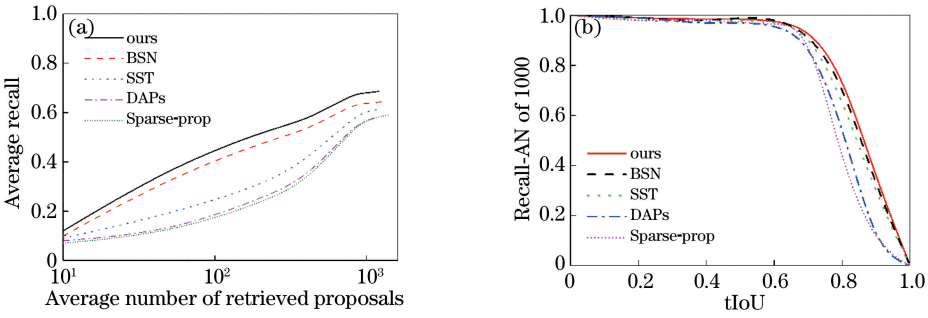

图 4　候选区域在 Thumos 14 上的表现。(a) AR-AN;(b) Recall-AN-tIoU

Fig. 4　Performances of candidate regions on Thumos 14. (a) AR-AN; (b) Recall-AN-tIoU

体两段式时序区域选举算法(Watershed+Reg)产生的候选区域的质量分别进行评估,见表 1。

表 1　不同时序选举方法在 Thumos 14 数据集上产生的候选区域在后续定位任务中的表现

Table 1　Performances of different temporal proposal methods in subsequent localization in candidate regions on Thumos 14　%

Method	DAPs+SVM	SVM	SCNN localizer
DAPs[14]	13.9	9.5	16.3
Sparse-prop[15]	7.8	8.1	15.3
SST[16]		15.9	23.0
BSN[17]		20.7	29.4
SCNN-prop[10]	7.6	14.0	19.0
Watershed	4.9	6.7	15.2
Reg	8.4	9.9	18.6
Watershed+Reg	24.7	23.8	37.2

结果显示,不论后续连接 SVM 分类器还是 SCNN 定位器,两段式时序区域选举算法产生的候选区域的精度均高于其他算法。两段式时序区域选举算法产生的候选区域连接 SCNN 定位器后,定位精度比 DAPs[14]方法高出 20.9%,比 Sparse-prop 高出 21.9%,比 SST 方法高出 14.2%,比 BSN 方法高出 7.8%。其中 DAPs 利用 C3D 提取视频特征,BSN 利用全局化的边界敏感网络,SST 引入非常长的视频序列。注意到,相比于 SCNN[10]原始的三段式架构,将前两段 SCNN-prop 替换为“分水岭+时序回归”两段式选举算法后,准确率提升了 18.2%。验证了两段式区域选举算法产生的候选区域具有高质量特性,也说明高质量的时序区域选举算法对时序定位任务具有促进作用。

直接将第一段改进的分水岭选举算法“Watershed”得到的候选区域输入定位器,精度就能够达到 15.2%,而“Watershed+Reg”两段式组合的形式与其他方法相比则达到了最高的识别精度。这是因为第一段选举算法为第二段选举算法提供了具有动作可能性依据的时序区间,且考虑了动作长度的多样性,使得回归结果更加精准。这验证了第一段选举算法的有效性,也说明了两段式选举算法的必要性。

ActivityNet v1.3 测试集的结果如表 2 所示,评估方法采用 ActivityNet 官方评估包。将两段式候选区域选举算法与 SCNN[10]三段式架构的第三段相连,可以看到,与其他近期时序定位算法 BSN、CDC、TCN、SCC 相比,本文算法取得了很高的平均精度均值,并且在很高的阈值(0.75 和 0.95)下仍能取得较高的定位精度。

表 2　ActivityNet v1.3 数据集中,各方法时序定位坐标在不同 tIoU 下的 mAP

Table 2　mAP of temporal localization coordinate of each method with different tIoU on ActivityNet v1.3 dataset　%

tIoU	0.5	0.75	0.95	Average
Method in Ref. [3]	42.28	3.76	0.05	14.85
BSN[17]	46.45	29.96	8.02	30.03
CDC[18]	45.30	26	0.20	23.80
TCN[19]				23.58
SCC[20]	40	17.90	4.70	21.70
Ours	48.58	31.74	8.71	31.23

其中:CDC[18]也运用了 C3D,在其结构上进行卷积和反卷积的改进,并且同样对结果进行了微调;BSN[17]利用全局化的边界敏感网络,同样提取了动作候选区域的特征,并且加入了起始和结束边界特征;TCN[19]同样引入动作的时序上下文信息,并且利用成对尺度采样特征对候选区域进行排序。这强有力地证明两段式候选区域选举算法的高精度性和有效性。

两段式候选区域选举算法产生的时序区间如图 5 所示。GT 代表真值坐标,Watershed 代表只用分水岭选举法得到的初级候选区域,Watershed+Reg 代表结合时序回归的两段式区域选举算法生成的候选区域。

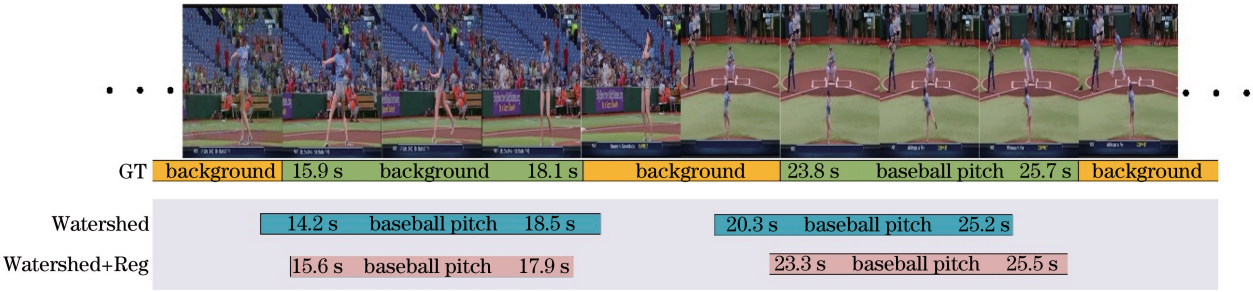

图 5　两段式区域选举算法在各阶段产生的候选区域示意图

Fig. 5　Schematic of candidate regions generated in each stage of two-stage regional proposal algorithm

4.3　消融研究

为验证本文时序上下文信息模块、时序金字塔结构化模块、C3D 单元级别特征训练这 3 个模块的有效性和必要性，对两段式时序区域选举网络的 4 种变体进行比较。1) w/o ctx & w/o pyramid：去掉时序上下文扩展模块和时序金字塔结构化模块，w 为 with 的缩写，代表包含某模块，w/o 为 without 的缩写，代表不包含某模块；2) w/ctx & w/o pyramid：保留时序上下文扩展模块，但不对内部候选区域进行特殊的 $l_{\text{level}}=2$ 的金字塔池化；3) frame reg w/both：保留上述两个模块，但去掉 C3D 单元级别特征提取模块，直接将帧图片输入到网络中进行帧级别定位；4) unit reg w/both：保留所有模块。通过 AR-AN 图来对比这 4 种变体的优劣，如图 6 所示。

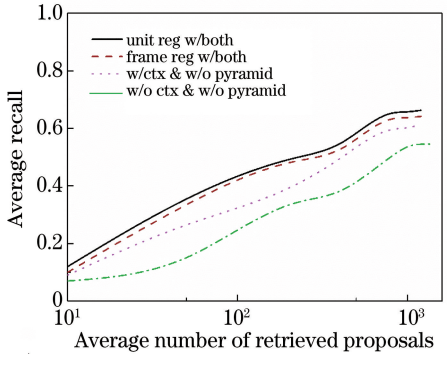

图 6　时序上下文信息模块、时序金字塔模块、C3D 单元级别特征的消融研究

Fig. 6　Ablation study oftemporal context module, temporal pyramid module, and C3D unit-level feature

可以看出，时序上下文扩展模块为时序检测任务提供了额外的语义信息，增强了候选区域的动作完整性。因此包含时序上下文模块的网络 w/ctx & w/o pyramid 比 w/o ctx & w/o pyramid 检测精度更高。没有针对内部候选区域实施金字塔结构化处理的 w/o ctx & w/o pyramid 识别精度不如运用金字塔结构化处理的 frame reg w/both，说明时序金字塔结构化模块的引入可以更具针对性地对动作本身（内部候选区域）及其上下文信息（扩展候选区域）建模，提升了准确度。而运用 C3D 单元级别特征的 unit reg w/both 比只利用帧级信息的网络 frame reg w/both 精度更高，说明挖掘了视频时域和空域信息的 C3D 单元级别特征可以提升时序区间定位的准确度。

4.4　效率分析

为验证两段式区域选举算法的高效性，保持 4.1 节参数设置不变，将两段式区域选举算法的 3 种变体（Proposal-16，Proposal-32，Proposal-w/o unit）与其他方法进行运行速度（FPS）和召回率（AR-AN of 1000）上的比较。其中 Proposal-16 和 Proposal-32 分别代表以一个单元为 16 frame 或 32 frame 提取的单元级 C3D 特征，Proposal-w/o unit 代表不使用单元级特征，只用单帧特征进行时序定位。随机抽取 Thumos 14 验证集中的 100 个视频，并在一个单个的 Nvidia TITAN X GPU 上进行训练，运行结果如表 3 所示。

表 3　各方法运行速度和召回率在 Thumos 14 数据集上的比较

Table 3　Comparison of FPS and recall rate of different methods on Thumos 14 dataset

Method	AR-AN of 1000	FPS
DAPs[4]	57.64	134.30
Sparse-prop[2]	56.60	10.20
SST[3]	60.27	308
CDC[18]		500
Proposal-16	66.27	423.15
Proposal-32	62.35	760.84
Proposal-w/o unit	60.41	129.40

可以看出：精度上 Proposal-16、Proposal-32、Proposal-w/o unit 依次降低；速度上，Proposal-32、Proposal-16、Proposal-w/o unit 依次降低；

Proposal-w/o unit 在速度和精度上都是最低的。这说明 C3D 单元级别特征的引入从整体上提升了算法的效率和精度。此外，更长的单元级别特征速度更快（Proposal-32 的速度高于 Proposal-16），但同时降低了召回率的准确度（Proposal-32 的召回率小于 Proposal-16）。AR 和 FPS 之间呈现一种此消彼长的关系，这是由于更短的单元级别特征会产生更多的特征向量，从而消耗更多的计算时间，但由于回归的基础坐标单位变小，更短的单元级别特征会提升时序坐标回归时的精确度。与其他时序动作选举方法相比，Proposal-16 超越了 DAPs、Sparse-prop 和 SST，达到了最优的召回率，并且速度比传统的 DAPs 方法提升了 3 倍。其中 CDC 同样利用动作边界进行微调，后续应用 C3D，但速度慢于 Proposal-32，这是因为相比于传统的手工特征或传统的单帧训练，单元级别特征的引入提升了训练效率。由表 2 可知，CDC 的平均精度均值（23.80%）远低于 Proposal-16（31.23%）。因此选择 C3D 单元级别特征（n_u=16）作为时序回归网络的输入，在大幅度提升时序定位精度的同时，保证了相对较高的处理速度。

5　结　论

针对长视频中时序动作发生时间的定位任务，提出一种两段式时序区域选举算法。第一段选举算法通过改进的分水岭浸水聚类产生多种不同长度的候选区域，实现动作时序边界的粗定位。进而提出一种时序金字塔结构化分析方法，引入时序上下文信息模块，对候选区域进行结构化建模，增强了动作片段的结构特征。第二段时序回归选举网络采用多任务损失函数，实现时序坐标回归器和动作/背景分类器的同时训练，过滤掉背景区域的同时得到了更加精确的时序边界。整个两段式区域选举网络以 C3D 提取的单元级视频特征进行训练，挖掘了视频时域和空域的丰富语义，提升了算法的精度和训练效率。多组实验结果表明，结合分水岭和回归网络的时序动作选举算法能够有效提高视频时序动作检测的准确度。

参　考　文　献

[1] Oneata D, Verbeek J, Schmid C. The LEAR submission at thumos 2014[M] // Fleet D, Pajdla T, Schiele B, *et al*. European conference on computer vision-ECCV 2014. Lecture notes in computer science Cham, 2014, 8692: 1-7.

[2] Li Y D, Xu X P. Humanaction recognition by decision-making level fusion based on spatial-temporal features[J]. Acta Optica Sinica, 2018, 38(8): 0810001.
李艳荻, 徐熙平. 基于空-时域特征决策级融合的人体行为识别算法[J]. 光学学报, 2018, 38(8): 0810001.

[3] Li Q H, Li A H, Wang T, *et al*. Double-stream convolutional networks with sequential optical flow image for action recognition[J]. Acta Optica Sinica, 2018, 38(6): 0615002.
李庆辉, 李艾华, 王涛, 等. 结合有序光流图和双流卷积网络的行为识别[J]. 光学学报, 2018, 38(6): 0615002.

[4] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 4489-4497.

[5] Gorban A, Idrees H, Jiang Y G, *et al*. THUMOS challenge: action recognition with a large number of classes[OL]. 2015[2019-05-25]. http: // www.thumos.info/.

[6] Feng X Y, Mei W, Hu D S. Aerial Target detection based on improved Faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[7] Xin P, Xu Y L, Tang H, *et al*. Fast airplane detection based on multi-layer feature fusion of fully convolutional networks[J]. Acta Optica Sinica, 2018, 38(3): 0315003.
辛鹏, 许悦雷, 唐红, 等. 全卷积网络多层特征融合的飞机快速检测[J]. 光学学报, 2018, 38(3): 0315003.

[8] Felzenszwalb P F, Girshick R B, McAllester D, *et al*. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[9] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C] // 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), June 17-22, 2006, New York, NY, USA. New York: IEEE, 2006.

[10] Shou Z, Wang D, Chang S F. Temporal action localization in untrimmed videos via multi-stage

CNNs[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 1049-1058.

[11] Donahue J, Hendricks L A, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 2625-2634.

[12] Wang L M, Qiao Y, Tang X O, *et al*. Actionness estimation using hybrid fully convolutional networks [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2708-2717.

[13] Wang L M, Xiong Y J, Wang Z, *et al*. Temporal segment networks: towards good practices for deep action recognition[M] // Leibe B, Matas J, Sebe N, *et al*. European conference on computer vision-ECCV 2016. lecture notes in computer science. Cham: Springer, 2016, 9912: 20-36.

[14] Escorcia V, Caba Heilbron F, Niebles J C, *et al*. DAPs: deep action proposals for action understanding [M] // Leibe B, Matas J, Sebe N, *et al*. European conference on computer vision-ECCV 2016. lecture notes in computer science. Cham: Springer, 2016, 9907: 768-784.

[15] Heilbron F C, Niebles J C, Ghanem B. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 1914-1923.

[16] Buch S, Escorcia V, Shen C Q, *et al*. SST: single-stream temporal action proposals[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6373-6382.

[17] Lin T W, Zhao X, Su H S, *et al*. BSN: boundary sensitive network for temporal action proposal generation[M] // Ferrari V, Hebert M, Sminchisescu C, *et al*. European conference on computer vision-ECCV 2018. lecture notes in computer science. Cham: Springer, 2018, 11208: 3-21.

[18] Shou Z, Chan J, Zareian A, *et al*. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1417-1426.

[19] Dai X Y, Singh B, Zhang G Y, *et al*. Temporal context network for activity localization in videos[C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 5727-5736.

[20] Heilbron F C, Barrios W, Escorcia V, *et al*. SCC: semantic context cascade for efficient action detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 3175-3184.