软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(5):1865−1879 [doi: 10.13328/j.cnki.jos.006202]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

Grenander 时间结构学习与推理优化下的行为识别*

吴克伟[1,2,3], 高 涛[1,2,3], 谢 昭[1,2,3], 郭文斌[1,2,3]

[1](大数据知识工程教育部重点实验室 (合肥工业大学), 安徽 合肥 230601)

[2](情感计算与先进智能机器安徽省重点实验室 (合肥工业大学), 安徽 合肥 230601)

[3](合肥工业大学 计算机与信息学院, 安徽 合肥 230601)

通信作者: 谢昭, E-mail: xiezhao@hfut.edu.cn

摘 要: 针对现有基于视频整体时间结构建模的行为识别方法中, 存在的时间噪声信息和歧义信息干扰现象, 从而引起行为类别识别错误的问题, 提出一种新型的 Grenander 推理优化下时间图模型 (temporal graph model with Grenander inference, TGM-GI). 首先, 构建 3D CNN-LSTM 模块, 其中 3D CNN 用于行为的动态特征提取, LSTM 模块用于该特征的时间依赖关系优化. 其次, 在深度模块基础上, 利用 Grenander 理论构建了行为识别的时间图模型, 并设计了两个模块分别处理慢行为时间冗余和异常行为干扰问题, 实现了时间噪声抑制下的时间结构提议. 随后, 设计融合特征约束和语义约束的 Grenander 测度, 并提出一种时序增量形式的 Viterbi 算法, 修正了行为时间模式中的歧义信息. 最后, 采用基于动态时间规划的模式匹配方法, 完成了基于时间模式的行为识别任务. 在 UCF101 和 Olympic Sports 两个公认数据集上, 与现有多种基于深度学习的行为识别方法进行比较, 该方法获得了最好的行为识别正确率. 该方法优于基准的 3D CNN-LSTM 方法, 在 UCF101 数据集上识别精度提高 6.41%, 在 Olympic Sports 数据集上识别精度提高 5.67%.

关键词: 行为识别; 时间模式; Grenander 时间图模型; 深度模型; 动态时间规划

中图法分类号: TP181

中文引用格式: 吴克伟, 高涛, 谢昭, 郭文斌. Grenander时间结构学习与推理优化下的行为识别. 软件学报, 2022, 33(5): 1865–1879. http://www.jos.org.cn/1000-9825/6202.htm

英文引用格式: Wu KW, Gao T, Xie Z, Guo WB. Temporal Structure Learning with Grenander Inference for Action Recognition. Ruan Jian Xue Bao/Journal of Software, 2022, 33(5): 1865–1879 (in Chinese). http://www.jos.org.cn/1000-9825/6202.htm

Temporal Structure Learning with Grenander Inference for Action Recognition

WU Ke-Wei[1,2,3], GAO Tao[1,2,3], XIE Zhao[1,2,3], GUO Wen-Bin[1,2,3]

[1](Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230601, China)

[2](Anhui Province Key Laboratory of Affective Computing & Advanced Intelligent Machine (Hefei University of Technology), Hefei 230601, China)

[3](School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

Abstract: Action recognition is one crucial and very challenging task in computer vision. Most of the existing methods use the temporal structure of the whole video and ignore its temporal noise and ambiguity feature, which leads to failure in action recognition. To address this problem, a novel temporal graph model is proposed with Grenander inference, namely, TGM-GI. First, a 3D CNN+ LSTM module is constructed to learn deep features, in which 3D CNN extracts the dynamic feature of video clips and LSTM optimizes the time dependence between features of two clips. Second, a temporal graph model is constructed with these deep features which use the generator

* 基金项目: 国家重点研发计划 (2017YFB1002203); 国家自然科学基金 (61503111); 安徽省自然科学基金 (1808085MF168); 中央高校基本科研业务费专项资金 (PA2020GDSK0059)

收稿时间: 2020-05-08; 修改时间: 2020-06-27, 2020-09-16; 采用时间: 2020-11-02

space of Grenander theory. The original temporal pattern is modified using two operations, in which combination operation can remove redundancy clips like slow motion and denoise operation can remove low-frequency clips like abnormal motion. Third, an incremental Viterbi algorithm is proposed for temporal pattern learning with Grenander inference, in which a Grenander measure is designed with both feature bond and semantic bond. Finally, the dynamic time warping is used to match the Grenander temporal pattern of test video with the Grenander temporal pattern of the training set and the label of the test video is predicted. The experimental results show that the proposed TGM-GI outperforms the state-of-the-art methods on two acknowledge databases. The TGM-GI is superior to the baseline method of 3D CNN-LSTM, and its accuracy improves 6.41% on the UCF101 dataset and 5.67% on the Olympic Sports dataset respectively.

Key words: action recognition; temporal pattern; Grenander's temporal graph model; deep model; dynamic time warping

1 引 言

行为识别是计算机视觉的重要任务之一, 正被广泛应用于视频监控, 医学看护, 视频检索, 视觉问答, 人机交互, 虚拟现实等领域[1,2]. 行为识别的目的不仅是预测行为标签, 而且还需要预测行为的时间结构. 深度学习为行为识别提供了有效的表示方式, 但是, 由于在真实世界中的行为视频中, 存在时间噪声信息和歧义信息干扰, 行为识别仍然是一个对输入数据敏感的病态问题.

深度学习将行为识别视为多值分类问题. 其研究方向之一是采用卷积神经网络 (convolutional neural network, CNN) 来学习行为的深度表达, 主要包括 3 类模型, 2D CNN 用于描述 RGB 视频帧中的静态特征, 双流 CNN 在 RGB 通道基础上通过引入光流算子来增强运动特征描述, 3D CNN 用于描述 RGB 视频段中的动态特征. 融合 2D CNN 和膨胀操作的 I3D 模型[3]和 3D CNN 结构搜索模型[4], 已经成功应用于行为识别, 获得了领先的行为识别性能. 深度学习的另一个研究方向是循环神经网络 (recurrent neural network, RNN), 其典型代表是 LSTM 通过构建行为的时间结构, 以约束视频帧/段的表达[5]. 由于真实视频中存在的大量噪声干扰, 采用时间融合和时间注意力的方法已经成为该领域重要的研究方向. Tran 等人通过设计分组卷积结构来实现时间特征融合[6]. Wang 等人在 2D CNN 基础上, 关注于融合注意力的稀疏时间采样方法[7]. 然而, 上述深度模型仅使用隐式的网络结构来进行时间结构搜索, 没有设计显式的时间结构表达, 对时间模式的解释能力不强. 因此, 现有研究转向设计时间图结构模型, 来实现行为的时间结构优化. Kukleva 等人在手工特征基础上, 采用隐 Markov 模型来描述行为的时间结构[8]. De Souza 等人在手工特征基础上, 采用 Grenander 理论构建行为时间图模型[9]. De Souza 等人认为视频时间结构的合理性关键在于视频段之间的约束, 因此, 通过设计 Grenander 测度来描述时间结构约束, 并采用 MCMC 模拟退火方法来实现时间结构搜索.

然而, 现有的时间结构学习存在以下 3 个主要缺点. (1) 现有的隐式时间注意力深度方法[6,7], 采用固定粒度的视频帧/段, 其时间结构的长度是固定的, 无法鲁棒处理不同长度的时间序列, 以及其中存在的时间噪声信息. (2) 现有的显式时间图模型方法[8,9], 仍然采用手工特征, 该特征难以描述复杂行为现象. 同时, 视频中的时间噪声信息, 进一步干扰了基于手工特征的时间节点的语义判断. (3) 现有的 MCMC 模拟退火方法[9], 在搜索时间结构上的时间复杂度较高, 导致 Grenander 理论很难解决多行为类别多时间节点状态的语义推理任务.

为了解决上述问题, 本文将行为识别任务拆分为时间模式学习和时间模式匹配两个子任务, 提出了一种 Grenander 推理优化下时间图模型 (temporal graph model with Grenander inference, TGM-GI), 来实现行为的时间结构学习和行为识别任务. 本文采用时间图模型实现时间模式学习, 并将其中的时间图结构优化问题, 定义为一种二阶段模型. 在第 1 阶段, 本文设计了基于 3D CNN-LSTM 模型实现深度特征提取. 在第 2 个阶段, 本文在 Grenander 模式理论下, 设计 3 个新模块来解决行为的时间模式学习中的慢行为冗余、异常干扰、歧义信息问题. 在获得可靠的行为时间模式基础上, 本文采用基于动态时间规划的模式匹配方法, 最终完成行为识别任务. 本文的主要贡献如下.

(1) 本文首次对时间图模型使用 Grenander 模式理论进行图结构优化, 提出一种 Grenander 推理优化下时间图模型. 该模型将行为识别任务中的时间图模型的图结构优化问题, 定义为一种二阶段模型. 在第 1 阶段, 本文设计了基于 3D CNN-LSTM 模型实现深度特征提取. 在第 2 阶段, 本文设计了 Grenander 模式理论构建时序图模型, 并

实现行为模式的图结构优化.

(2) 本文的 Grenander 模式推理关注于行为的连续时间模式学习. 本文设计相似性合并模块和低概率特征抑制模块, 分别用于去除同质冗余和异常干扰. 同时, 本文设计了一种时序增量形式的 Viterbi 算法, 来实现对 Grenander 时间模式的优化, 解决二义性视觉干扰问题.

(3) 在 UCF101 和 Olympic Sports 两个公认数据集上, 与现有多种基于深度学习的行为识别方法进行比较, 本文方法获得了最好的行为识别正确率. 本文方法优于基准的 3D CNN-LSTM 方法, 在 UCF101 数据集上提高 6.41%, 在 Olympic Sports 数据集上提高 5.67%. 同时, 多种消融实验能够有效说明本文方法在行为的时间模式学习和行为识别上的可靠性.

2 现 状

早期的行为识别方法, 通常采用光流特征和轨迹特征实现行为表示[10,11], 在这些手工特征基础上的字典学习表示可以进一步获得行为原子的表示[12,13]. 然而, 近年来, 深度学习已经取代了手工特征方法. 本文主要阐述 3 种基本结构的研究现状, 卷积神经网络 CNN, 循环神经网络 RNN 以及图模型.

2D CNN 通过堆叠卷积网络来学习图像中复杂的视觉模式. 与图像分类中的空间模式不同, 行为识别中还需要提取视频中的时空模式. 双流 CNN 通过引入光流, 来增强运动特征提取能力[14,15]. Zhu 等人[16] 认为传统的光流算子不足以提取复杂运动特征, 并提出了一种隐双流网络, 即隐式的卷积网络来隐式地学习类似光流的特征. 另一种更直接的时空特征提取方法是 3D CNN 网络[17,18], 通过对输入的多帧构成的视频段来进行时空模式学习. 3D CNN 仍然是目前领先的行为识别方法[1,2]. Carreira 等人[3]通过对 2D CNN 的参数进行膨胀获得 3D CNN 网络结构, 有效提高了时空特征提取的准确性. Tran 等人[4]将 3D CNN 拆分为 R(2+1)D 对空间特征和时间特征进行分离学习. Xiao 等人[19]构建了基于三级空间金字塔的 3D CNN 网络. Tran 等人[20]构建了融合残差模块的 3D CNN 网络. 此外, 3D CNN 的设计还引入了不同的线索, 包括, 空间注意力[21], 高层语义特征[22]和类似光流特征[23]. 但是, 上述方法都是以固定粒度的方式进行视频帧/段的采样, 从而无法抑制视频中存在的冗余信息和不相关干扰, 限制了行为识别的准确性.

基于 CNN 行为识别的另一个重要研究方向是时间结构优化. 时间特征的加权池化是一种有效的时间结构优化方法. Lin 等人[24]设计了时间平移模块, 用于融合视频帧之间的特征. Zolfaghari 等人[25]设计 2D 与 3D 的并行 CNN 结构来融合视频中的长期特征. Tran 等人[6]通过设计分组卷积结构来实现时间特征融合. 优化行为时间结构的另一种线索是时间注意力估计[26]. Wang 等人[7]采用时间注意力加权, 来实现多尺度时间窗的池化. Wang 等人[27]采用视频段的重要性排序, 实现时间结构选择. Long 等人[28]设计了融合时间注意力的 CNN 模型. 现有方法还研究了时间结构中的时间关系表示和推理[29]. Xu 等人[30]使用自监督学习方法, 来估计时间顺序关系. Wu 等人[31]使用长时间特征池方法, 来估计特征之间的时间依赖关系. Dwibedi 等人[32]学习时间中的周期性依赖, 来约束时间模式学习. 虽然, 上述研究已经关注于时间结构的学习, 但是, 并没有给出显式的时间结构来解释行为的执行过程, 仍然难以充分抑制行为的时间干扰信息.

与 CNN 的视频帧/段的特征提取不同, 循环神经网络 RNN 关注于时间依赖关系建模, 并可以通过反向传播算法对时间依赖进行优化[33]. LSTM 作为一种特殊的 RNN 模型[34], 由于具有学习长期依赖关系的能力, 受到广泛的关注. Ng 等人[35]研究了 LSTM 中的时间特征池结构设计. Li 等人[5]设计了融合运动注意力的 LSTM 模型. 为了更好地对行为特征进行空间和时间建模, CNN-LSTM 方法已经受到研究者的广泛关注, 其中 Ullah 等人[36]使用 2D CNN 作为 LSTM 的输入特征, Ouyang 等人[37]考虑 3D CNN 与 LSTM 结合, Song 等人[38]进一步讨论在该框架下, 2D CNN 特征和 3D CNN 特征的互补性. 由于上述模型仍然是 RNN 模型的变形, 其对时间结构建模是隐式的, 仅能预测视频级的行为标签, 而不能给出视频帧/段级的标签, 因此, RNN 仍然不能满足人类对行为时间结构的理解要求.

时间图模型更符合人类对行为的自然理解, 有助于时间结构的优化. 早期的时间图模型使用手工特征学习行为原子[39], 并通过聚类方法[40]和字典学习方法[41]实现对视频帧/段级的行为标签预测. Lan 等人[40]实现了一种层

级结构视频表示方法, 并使用聚类方法发现多种粒度下的中层行为元素. 隐 Markov 模型 (hidden Markov model, HMM) 使用概率图模型来描述时间结构, 并可以通过时间递归的方式对行为原子进行链接, 获得长时间行为时间结构[41], 以描述行为的时间复合状态. Kuehne 等人[42]在行为原子的基础上, 学习行为状态的语法描述. Li 等人[43]在视觉单词基础上, 使用二进制动态模型来描述行为的复合状态. Kukleva 等人[8]使用 Viterbi 算法可以对 HMM 框架中的时间结构进行有效地优化. Grenander 理论针对 HMM 的隐状态空间和时间结构约束进行改进, 可以有效解释视频帧/段级的语义内容[9]. De Souza 等人[9]采用 MCMC 模拟退火方法实现的基于 Grenander 理论的行为状态解码. 然而, 上述图模型方法虽然实现了一定的时间图模型构建, 但是, 他们均是采用手工特征, 不能满足真实视频的行为识别内容复杂性, 也没有设计可变长度的时间结构, 难以满足真实视频的时间结构复杂性的需要. 同时, MCMC 模拟退火方法在多行为类别多节点状态的推理时间复杂度较高, 也直接降低了 Grenander 理论在行为识别中的应用能力.

为了解决上述问题, 本文首次对时间图模型使用 Grenander 模式理论进行图结构优化, 提出一种 Grenander 推理优化下时间图模型 (TGM-GI). 该模型将行为识别任务中的时间图模型的图结构优化问题, 定义为一种二阶段模型. 在第 1 阶段, 本文设计了基于 3D CNN-LSTM 模型实现深度特征提取. 在第 2 阶段, 本文设计了 Grenander 模式理论构建时序图模型, 并实现行为模式的图结构优化.

3 方 法

本文将行为识别任务拆分为时间模式学习和时间模式匹配两个子任务, 并采用时间图模型实现时间模式学习, 将其中的时间图结构优化问题, 定义为一种二阶段模型, 如图 1 所示. 在第 1 阶段, 本文设计了基于 3D CNN-LSTM 模块实现行为特征提取. 在第 2 阶段, 本文在 Grenander 模式理论下, 设计 3 个模块来解决行为的时间图模式推理问题. (1) 首先, 为了构建时序图模型, 本文在深度特征空间中, 采用 K-means 方法构建 Grenander 生成器空间, 用于初始化时间模式表示. (2) 其次, 为了抑制时序图模型中的同质冗余和异常干扰, 设计了相似性合并模块和低概率特征抑制模块, 用于时间模式提议. (3) 最后, 本文设计融合特征约束和语义约束的 Grenander 测度用于评价时间模式, 并构建了一种时间增量形式的 Viterbi 算法进行时间图模式推理, 来解决二义性视觉干扰问题, 以获得可靠的行为时间模式. 在获得可靠的行为时间模式基础上, 本文采用基于动态时间规划的模式匹配方法, 最终完成行为识别任务.

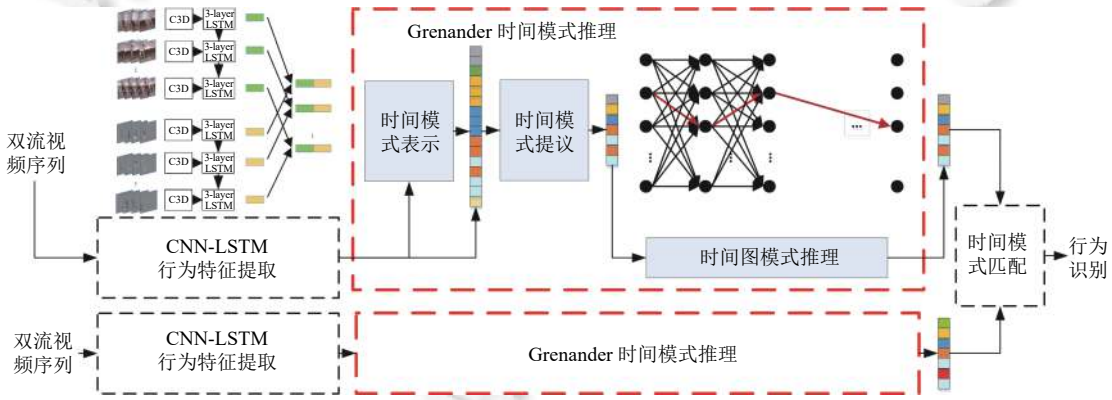


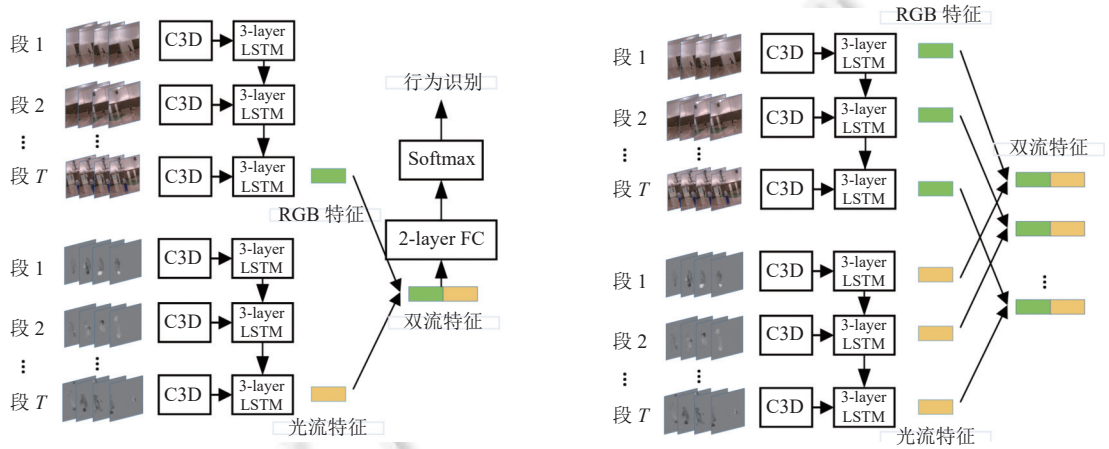
图 1 基于 Grenander 时间图模式推理的时间模式学习和行为识别

3.1 3D CNN-LSTM 深度网络结构

本文设计了 3D CNN-LSTM 模型来实现行为特征提取, 其中 3D CNN 用于提取视频段的动态特征, LSTM[34]用于进一步学习这些特征之间的时间依赖关系, 如图 2(a) 所示. 首先, 视频被等间隔划分为 16 段, 每段等间隔采样 16 帧. 每段的 16 帧输入到 C3D 网络结构中[18], 并提取该网络的 FC6 层特征作为该视频段的运动特征. 3D

CNN 比 2D CNN 能更好的描述视频段中的动态信息.

为了进一步分析 16 个视频段之间的时间依赖关系, 16 段的特征被输入到具有长度为 16 的 LSTM 模型中, 每个时间节点具有 3 个隐层单元的 LSTM 单元, 每个隐层单元的输出特征维度为 512 维. 本文进一步设计了双流的 C3D 和 3-layer LSTM 模型, 并将最后一个时间节点的 RGB 特征和光流特征串联, 输入到 2 层 FC 网络和 Softmax 层实现行为识别. 我们将图 2(a) 的双流 3D CNN-LSTM 模型作为本文的基准模型, 通过进一步添加推理模块, 可用于验证 Grenander 时间图模式推理的优势. 该基准模型采用交叉熵形式的损失函数, 进行端对端的方式学习, 以获得每个时间节点的特征. 对双流时间节点特征, 采用串联方式获得每个视频段的双流特征, 如图 2(b) 所示. 该双流特征构建了 Grenander 生成器的特征空间, 用于提取时间节点的隐语义标记.



(a) 双流的 CNN-LSTM 行为识别 (b) 双流的 CNN-LSTM 行为特征提取

图 2 双流的 CNN-LSTM 行为特征提取

3.2 基于 Grenander 理论的图结构搜索

本文采用 Grenander 理论的四元组 $R\langle G,S,\rho,\sum\rangle$ 来描述时间图模式 [9], 其中, G 是时间模式的生成器空间, 用于描述时间节点的隐语义; S 是生成器空间的划分, 同一个划分中的特征相似, 隐语义标记相同; ρ 是生成器之间的时间约束, 时间约束可以多个连接生成器, 以生成长时间的时间模式; $\sum$ 是时间图模式的配置空间, 包含所有可能的时间图模式.

3.2.1 Grenander 生成器空间和约束测度

构建 Grenander 图模型的第一步是构建生成器空间, 即对时间节点进行隐语义标记. 本文使用 K-means 方法对 3D CNN-LSTM 提取的特征空间进行聚类, 每个聚类 $m_1,m_2,\ldots,m_K$ 是一个生成器空间划分, 描述一类行为原子, 可以对每个时间节点进行隐语义标记. 通过调整 K-means 方法中的聚类数 K, 来改变隐语义标记的数量. 图 3(a) 给出了生成器空间的实例 $g\in G$, 其中每个生成器中间的聚类编号为该时间节点的隐语义.

构建 Grenander 图模型的第 2 步是设计时间约束, 时间约束可以连接多个生成器形成长时间模式 (图 3(b)). 由于时间约束是有向关系, Grenander 理论采用 in-bond 来描述时间前驱 (图 3(a) 中时间节点的白色半圆接口), 记作 β'', out-bond 来描述时间后继 (图 3(a) 中时间节点的黑色半圆接口), 记作 β'. 在真实视频中, 如果两个隐语义 g_i 和 g_j 存在时间有向共生关系, 则进一步可以使用约束测度来度量这对时间约束的强度 $(\beta'(g_i),\beta''(g_j))$. 由于点对隐语义受到特征约束和共生约束. 因此, 我们设计了一种融合特征约束和共生约束的时间模式点对测度, 记作

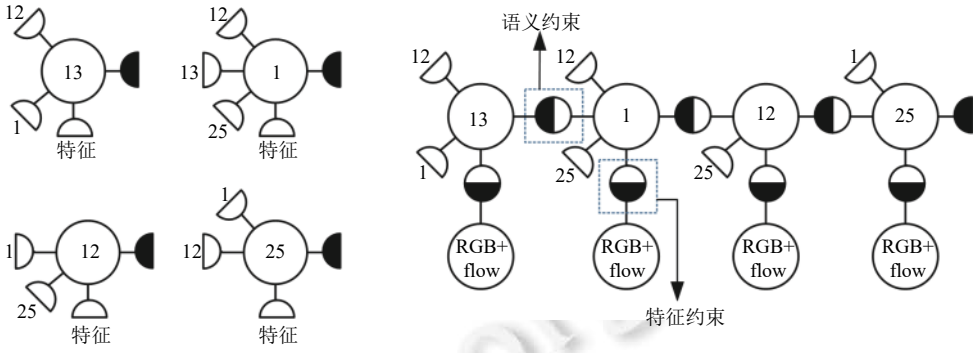
$$A(\beta'(g_i),\beta''(g_j))=\rho_s*\rho_g \tag{1}$$

其中特征约束 $\rho_g=\rho_{g_i}*\rho_{g_j}=e^{-\|m_i-h_i\|}*e^{-\|m_j-h_j\|}$, 是两个节点的相似度乘积, 每个节点的相似度使用每个节点特征和聚类特征的 2 范数距离, 并采用指数分布进行概率化. 其中共生约束 ρ_s, 是隐语义的转移频率矩阵, 通过对训练视频中的隐语义点对统计获得. 从图 3(b) 中可以看出, 整个时间模式是由序列的所有点对模式组成的, 其时间模式

的测度也通过点对模式测度的乘积获得.

3.2.2 Grenander 时间模式提议

利用聚类获得的生成器空间, 可以对 Grenander 时间模式进行隐语义标记, 获得初始时间模式 c^{init}, 该时间模式的时间节点配置可以记作 $c^{\text{init}}=\sigma(g_1,g_2,\ldots,g_T)$, 其中 $\sigma\in\Sigma$ 是该时间模式的时间节点配置, T 表示时间模式的长度.



(a) 生成器空间 (b) 约束和时间模式

图 3 Grenander 理论的生成器空间, 约束和时间模式

Grenander 时间模式学习的第 1 个步骤是优化时间结构, 去除其中可能存在的慢行为冗余和无关噪声干扰. 因此, 本文采用相邻隐语义标记合并的方法去除慢行为冗余. 本文使用低概率阈值参数 θ, 删除共生低概率的噪声时间节点, 即如果视频中的隐语义点对的共生概率小于阈值, 则去除其后继节点. 经过上述时间节点采样操作, 可以获得提议的时间模式 $c^{\text{prop}}=\sigma(g_1,g_2,\ldots,g_L)$, L 表示提议后的时间模式长度.

3.2.3 Grenander 时间模式推理

Grenander 时间模式学习的第 2 个步骤是优化隐语义标记. 初始的隐语义分配使用最近邻策略, 即将与节点特征最近的聚类中心标记, 记作当前节点的隐语义. 但是, 节点特征容易受到观测噪声影响, 噪声的干扰容易引的错误地隐语义标记, 尤其是当其处于多聚类中心边界处时, 因此, 需要考虑节点对之间的共生关系, 来平滑这些错误标记.

Grenander 时间模式推理, 是一种时间模式测度的最大化过程. 推理需要对时间模式进行可解释性的度量, 即先依次度量长时间模式中的点对关系, 随后乘积整个时间模式的所有点对关系, 以概率密度函数来描述其测度:

$$p(c)=\frac{1}{Z}\prod_{(\beta',\beta'')\in c}A(\beta'(g_i),\beta''(g_j)) \tag{2}$$

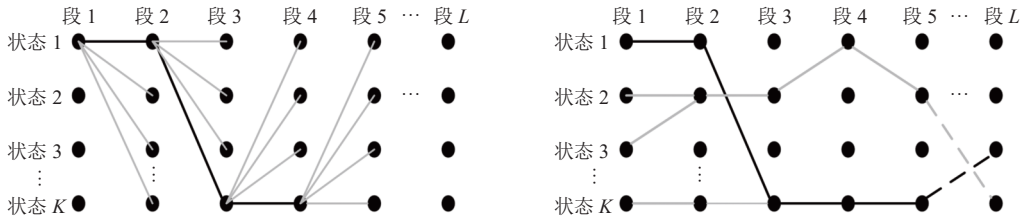
其中, Z 是归一化系数. 该概率密度越高说明隐语义状态序列越合理. 由于概率密度函数中的乘积操作容易造成精度损失, 进一步采用负对数函数 $E(c)=-\log p(c)Z$, 将乘积操作转化为加法操作, 同时, 将概率最大化求解, 转化为能量最小化求解, 其长时间模式的能量测度为:

$$E(c)=-\sum_{(\beta',\beta'')\in c}\log A(\beta'(g_i),\beta''(g_j)) \tag{3}$$

Grenander 推理, 需要在时间模式的状态空间中搜索到最小能量的隐语义标记. 然而, 该时间模式有 n 个时间节点, 每个时间节点有 K 个状态, 完全遍历一次需要 $O(K^n)$. 采用 MCMC 方法虽然能避免一定的局部最优解问题, 但是其搜索的时间复杂度也比较高 [9]. 因此, 本文在 Viterbi 算法基础上[8], 将 Grenander 推理视为一个多阶段决策过程, 并设计了一种增量形式的 Viterbi 算法依次推理出每个时间节点的标记. 这是因为, 我们假设最优隐状态中的每个时间前缀是最优的, 因此, 可以通过最优前缀结构减少搜索空间.

图 4(a) 给出了 Grenander 的增量推理形式, 推理从第 1 个节点开始依据时间依次遍历, 并对第 2 个节点的所有状态进行遍历 (图中灰色线表示遍历), 通过 Grenander 测度确定最优前缀子结构 (图中黑色线表示遍历), 随后依次确定后续节点的最优状态. 当遍历完所有时间节点后, 我们会得到 K 个不同初始状态的最优子结构 (图 4(b)), 在

评价候选子结构的 Grenander 测度, 获得最优的决策, 并最终完成 Grenander 推理. 算法 1 中进一步给出了 Grenander 时间模式推理的时间增量 Viterbi 算法, 由于只需要遍历 n 个时间节点, 每个节点遍历 K 个状态, 该算法的时间复杂度可以降低到 $O(K\cdot n)$.



(a) Grenander 时间模式的增量推理　　(b) Grenander 时间模式的最优决策

图 4　Grenander 时间模式的增量推理和最优决策

算法 1. Grenander 时间模式推理的增量 Viterbi 算法.

输入: 时间模式的长度 L, 时间模式中节点之间的特征相似度矩阵 ρ_g, 语义转移矩阵 ρ_s, 节点状态数量 K;
输出: Grenander 时间模式 c^g.

```
for k^1 = 1 : K // 遍历第一个节点的状态
    c_0^{start=k^1} = σ(k^1) // 初始化第一个节点的状态
    E_{0,1}^{start=k^1} = 0 // 初始化第一个节点的能量
    for t = 2 : L // 以时间增量方式, 确定后继节点的状态
        for k^t = 1 : K // 遍历后继节点的状态
            c_{t-1}^{start=k^1} = σ(k^t) // 获得前驱节点的状态
            E_{t-1,t}^{start=k^1}(k^t) = −log ρ_s(k^{t−1}, k^t) ρ_g(h_{t−1}, g_{t−1}) ρ_g(h_t, g_t) // 计算增量形式的 Grenander 测度
        end
        k^{t*} = arg min E_{t-1,t}^{start=k^1}(k^t) // 确定后继节点的状态
        c_{t-1}^{start=k^1} = σ(k^{t*}) // 记录后继节点的状态
        c_{0,1,...,t}^{start=k^1}(t−1) = c_{t-1}^{start=k^1} // 记录增量时间模式
        E_{0,1,⋯,t}^{start=k^1} = E_{0,1,⋯,t−1}^{start=k^1} + E_{t-1,t}^{start=k^1} // 更新增量时间模式的能量
    end
end
k^{0*} = arg min E_{0,1,...,L}^{start} // 确定 Grenander 时间模式
c^{inter} = c_{0,1,...,L}^{start=k^{0*}} // 记录 Grenander 时间模式
```

本文的图模式推理使用了 3 个新模块, 来解决时间模式中的慢行为冗余、异常干扰、歧义信息问题, 如图 5 所示. 首先, 通过 K-means 方法获得 Grenander 生成器空间, 以标记初始的时间模式 (图 5(a)); 随后, 对相同标记进行合并, 获得去除冗余的时间模式 (图 5(b)); 对低概率的时间标记对进行判断, 删除其后继节点, 获得去除异常的时间模式 (图 5(c)); 最后, 采用增量形式的 Grenander 推理, 修正其中的歧义标记, 获得 Grenander 时间模式 (图 5(d)). 因此, Grenander 时间模式, 对慢行为冗余、异常干扰、歧义信息具有较好的鲁棒性, 能够更好地描述时间模式, 用于区分不同的行为类别.

3.3　时间模式匹配

本文依次提取训练集合中各视频的 Grenander 时间模式, 构建行为类别的模板. 在测试过程中, 将测试视频的

Grenander 时间模式与训练获得的模板计算匹配相似度, 并将最大相似度的类别作为预测的行为类别. 我们采用动态时间规划的方法[44], 估计测试模式和训练模式的相似度, 是因为该算法允许跨时间匹配, 从而能够匹配长度不同的时间模式, 也进一步抑制模式中可能存在的噪声干扰.

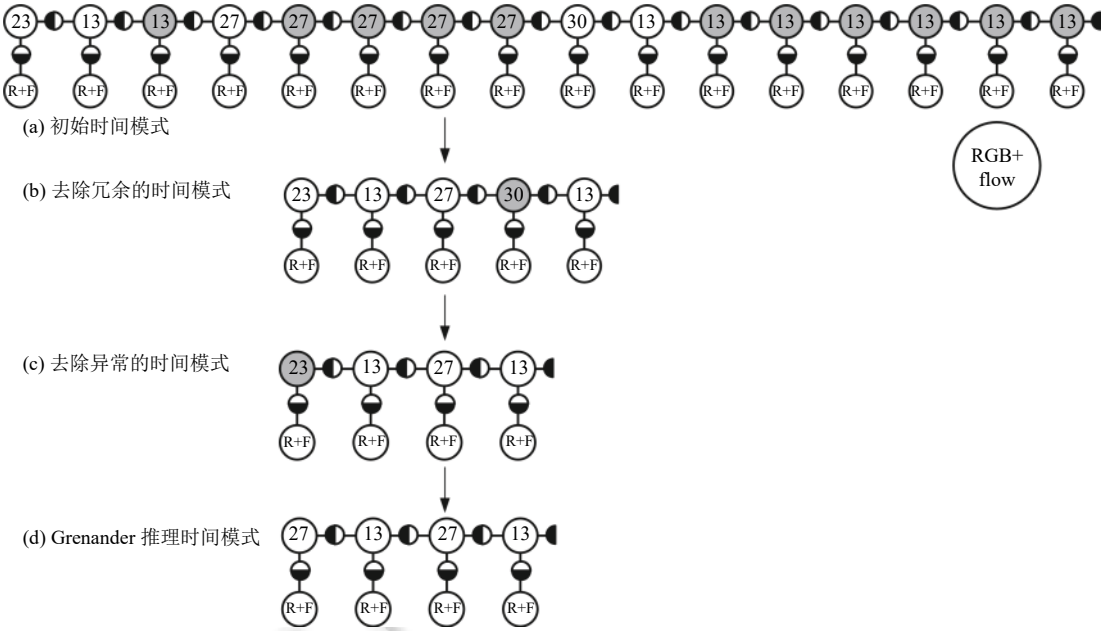


图 5 基于 Grenander 推理的时间模式学习过程

4 实 验

4.1 数据集

我们使用两个公认的行为识别数据库进行验证. UCF 101 数据库包含 101 个类别, 13 320 个视频, 每类动作至少包含 100 个视频. 该数据集中包含静态背景、观测视角变化、姿态等多种干扰. 我们将该数据集的 75% 视频用于训练, 25% 视频用于测试. Olympic Sports 数据集, 包括 16 类体育动作, 例如, 跳高, 撑杆跳高, 铁饼等, 共 783 个视频. 我们使用 649 个视频进行训练, 134 个视频用于测试.

4.2 实验细节

本文的双流 3D CNN-LSTM 模块, 采用 PyTorch 实现. 其中, 光流特征采用 TVL1[45]方法获得. 3D CNN 采用 C3D 网络架构, 并将其 FC6 的 4096 维特征输入到 LSTM. LSTM 采用 3 个隐层的 LSTM 模型, 每个隐层的输出维度是 512 维, 长度为 16 个时间节点. 深度网络损失函数采用交叉熵形式, 参数优化使用 Adam 方法, 其中学习率设置为 10^{-3}, 权重衰减为 0.0005. 批处理大小为 256, 训练迭代次数为 50. 实验需要条件为 2 块 Nvidia GeForce GTX 1080 显卡.

4.3 评价方法

本文使用平均分类正确率作为行为识别的指标, 即正确类别的视频数量与所有测试视频数据的比值.

4.4 定量对比

本文的对比方法包括 4 类, (1) 手工特征模型, 包括 IDT 特征方法[10], 运动正则特征方法[11], 谱特征方法[12], 特征堆叠方法[13]. (2) CNN 模型, 包括 I3D 网络[3], R(2+1)D 网络[4], 双流 CNN[15], 隐双流网络[16], C3D 网络[18], 带金字塔的 CNN[19], 双流 3Dnet[23], 带时间位移模块的网络 TSM[24], 融合外观和关系的网络 ARTNet[25], 时间分段网络

TSN[7]. (3) RNN 模型, LSTM 模型[5], CNN-LSTM 模型[36,37,38]. (4) 图模型, HMM[42], 结构化时间的 HMM[43]. 同时, 我们将本文设计的 3D CNN-LSTM 作为本文的基准方法, 表 1 和表 2 分别列除了本文方法在 UCF 101 和 Olympic Sports 上的实验结果, 说明本文方法能胜出所有对比方法.

表 1 本文方法与现有方法在 UCF101 上的平均分类正确率

方法分类	对比方法	基本结构	正确率 (%)
机器学习	Wang等人[10] (2013)	IDT, SVM	86.40
	Lan等人[13] (2015)	MIFS, SVM	88.60
深度学习(RGB)	Tran等人[18] (2015)	C3D	85.20
	Ouyang等人[37] (2019)	C3D+2 layer LSTM	88.90
	Song等人[38] (2018)	2D ResNet+3D ResNet+2 layer LSTM	91.10
	Ullah等人[36] (2018)	AlexNet + Bidirection 2 layer-LSTM	91.21
	Zolfaghari等人[25] (2018)	BN-Inception, 3D-ResNet18	94.80
	Carreira等人[3] (2017)	I3D	95.60
	Lin等人[24] (2019)	ResNet50	95.90
深度学习(隐双流)	Diba等人[23] (2016)	C3D	90.20
	Zhu 等人 [16] (2018)	I3D	97.10
深度学习(双流)	Simonyan等人[15] (2014)	ConvNet	88.00
	Li等人[5] (2018)	stacked LSTM	88.90
	CNN-LSTM [ours]	C3D, 3-layer LSTM	92.33
	Wang等人[27] (2017)	ResNet18	93.50
	Wang等人[7] (2019)	BN-Inception	94.90
	Tran等人[4] (2018)	R(2+1)D	97.30
	Carreira等人[3] (2017)	I3D	98.00
	TGM-GI [ours]	C3D, 3-layer LSTM	**98.74**

表 2 本文方法 Olympic Sports 上的平均分类正确率

方法分类	对比方法	基本结构	正确率 (%)
机器学习	Jones等人[12] (2014)	谱特征, SVM	74.60
	Kuehne等人[42] (2016)	HOG-HOF, HMM	90.20
	Wang等人[10] (2013)	IDT, SVM	91.10
	Ni等人[11] (2015)	FV特征, SVM	92.30
	Lan等人[13] (2015)	MIFS, SVM	92.90
	Li等人[43] (2017)	FV特征, HMM	93.10
深度学习(RGB)	Tran等人[18] (2015)	C3D	81.65
	Carreira等人[3] (2017)	I3D	83.74
	Xiao等人[19] (2019)	CNN	89.52
深度学习(双流)	3D CNN-LSTM [ours]	C3D, 3-layer LSTM	89.52
	TGM-GI [ours]	C3D, 3-layer LSTM	**95.19**

为了便于比较, 我们在表 1 和表 2 中, 进一步标记出机器学习方法使用的手工特征和机器学习模型, 深度学习方法使用的主干网络. 通过分析表 1 和表 2 的实验结果们可以看出, (1) 在 UCF101 数据库上, 虽然本文基准模型 3D CNN-LSTM, 没有现有的 I3D[3]准确率高, 但是, 利用 Grenander 理论增强后, 本文的 TGM-GI 方法胜出了现有最好的深度学习模型. (2) 本文的基准方法与 Ouyang X 方法[37], 都采用 C3D+LSTM 框架, 我们方法胜出的原因是, 我们使用了 3 layer LSTM 比 Ouyang X 采用的 2 layer LSTM 具有更好的非线性拟合能力. 此外, 我们的基准方法在 LSTM 后使用了 2 layer FC 来增强非线性表达, 而 Ouyang X 在 LSTM 之后, 使用的是 Mean pooling 和逻辑回

归, 其行为特征表达能力可能较弱. (3) 在 Olympic Sports 数据库上, 虽然 3D CNN-LSTM 的深度模型, 并不一定优于 HMM 方法[42], 同样, 利用 Grenander 理论增强后, 本文的 TGM-GI 方法胜出了所有图模型方法. (4) 部分深度学习在 Olympic Sports 数据集正确率较低, 其原因是, 由于拍摄视角多样性和背景干扰较多, 造成了 Olympic Sports 数据集存在噪声较多的情况, 这种情况下, 参数较多的深度学习方法容易陷入过拟合. Olympic Sports 训练样本较少, 进一步加剧了过拟合程度. (5) 本文的 TGM-GI 方法通过抑制时间噪声, 其结果优于基准的 3D CNN-LSTM 方法, 在 UCF101 数据集上提高 6.41%, 在 Olympic Sports 数据集上提高 5.67%. 因此, 充分说明了 Grenander 优化的时间图模式的有效性.

为了分析本文方法的计算效率, 我们使用 FLOPs 来估计模型运行所需要的计算量, FLOPs 全称为 floating point operations 指浮点运算数. 表 3 中给出了 C3D, LSTM 模块的参数量和 FLOPs. 从表中可以看出相对于原始的 C3D 模型, 添加 LSTM 模块后, 其计算量没有明显添加. 我们进一步给出测试时各阶段的运行时间. 测试时使用单张 Nvidia GeForce GTX 1080 显卡, i7-5960X 处理器. 在第 1 阶段, 执行一个视频的行为识别时, 深度学习模型需要考虑 16 个时间节点, 即 16 次 C3D-RGB-LSTM, 和 16 次 C3D-RGB-LSTM. 16 次 C3D-RGB-LSTM 的时间为 242 ms, 其中, C3D-RGB 部分 191 ms, LSTM 部分 51 ms. 16 次 C3D-FLOW-LSTM 的时间为 238 ms, 其中, C3D-FLOW 部分 186 ms, LSTM 部分 51 ms. 在第 2 阶段, 推理部分不仅需要考虑时间节点数量, 还需要考虑其状态数量, 其中, 初始隐语义标记、去除冗余部分和去除异常部分都需要一次遍历时间节点. 遍历 16 个节点的运行时间为 4×10^{-3} ms, 可以忽略不计. 此外, 我们测试了完全遍历的推理方法, 即暴力遍历每个时间节点的每个状态, 在时间状态数 K=160 情况下, 2 个时间节点需要时间 31 ms, 3 个时间节点需要时间 4 s, 4 个时间节点需要时间 657 s, 即 11 min, 这说明暴力遍历方法无法胜任实际应用需求, 也说明本文增量推理方法有效减少了运行时间.

表 3 本文模型的参数量和 FLOPs

深度模型	参数量	FLOPs
C3D-RGB	6.13E+07	7.70E+10
C3D-Flow	6.13E+07	7.64E+10
3-layer LSTM (16个时间节点)	2.52E+07	1.76E+08

4.5 消融实验

本文进一步设计消融实验来分析 Grenander 优化过程中的参数的影响, 主要包括 4 个关键参数, (1) 聚类数 K, 该参数影响时间节点的状态数; (2) 低概率阈值 θ, 该参数影响异常节点的去除, 数值越高会增加更多的异常节点; (3) Grenander 测度, 需要分别考虑语义约束和特征约束; (4) 时间模式的类型, 即图 5 中各阶段的时间模式.

(1) 聚类数的影响. 我们在公式 (1) 测度情况下, 来同时讨论聚类数 K 和低概率阈值 θ 的影响. 图 6 分别给出了 UCF101 和 Olympic Sports 数据库中, 本文方法 TGM-GI 在不同参数时的正确率. 可以看出聚类数 K 的影响如下: (1) UCF101 数据集上 K=160 时出现最高正确率, 而 Olympic Sports 数据集上 K=80 时出现最高正确率, 这是因为 UCF 具有更多的行为类别, 具有更多的隐语义行为原子, 需要更多的聚类数来对特征空间进行细分. (2) 但是随着行为原子的增加, 并不能提高正确率, 这可能是因为, 更多的行为原子之间难以相互区分, 反而导致了错误的时间模式标记.

(2) 低概率阈值的影响. Olympic Sports 取得最高正确率的条件为 K=80, θ=6; UCF101 取得最高正确率的条件为 K=160, θ=8. 我们注意到低概率阈值是相似的. 与聚类数 K 相比较, 低概率阈值的影响并不明显. 低概率阈值过低可能会导致模型退化而不删减异常节点, 该参数过高可能会导致出现次数较少但是具有判决能力的语义标记的删除, 这两种情况都会降低模型的正确率.

(4) Grenander 测度的影响. 表 4 进一步讨论 3 种 Grenander 测度的影响, 具体包括: (1) $A_1=\rho_s$ 只考虑语义约束; (2) $A_2=\rho_g$ 只考虑特征约束; (3) $A_3=\rho_s*\rho_g$ 同时考虑两种约束. 其他参数使用各数据集最高的参数条件, Olympic Sports 使用 K=80, θ=6, UCF101 使用 K=160, θ=8. 可以看出, 单语义约束的结果最差, 甚至比基准模型 3D CNN-LSTM 模型还低, 这可能是因为当出现了共生概率更高的语义节点对时, 语义标记被错误的分配, 反而导致错误的

时间模式标记. 为了避免这种情况, 在特征约束的基础上, 添加语义约束, 就可以同时兼顾特征相似度和共生频率, 从而获得更好的 Grenander 时间模式.

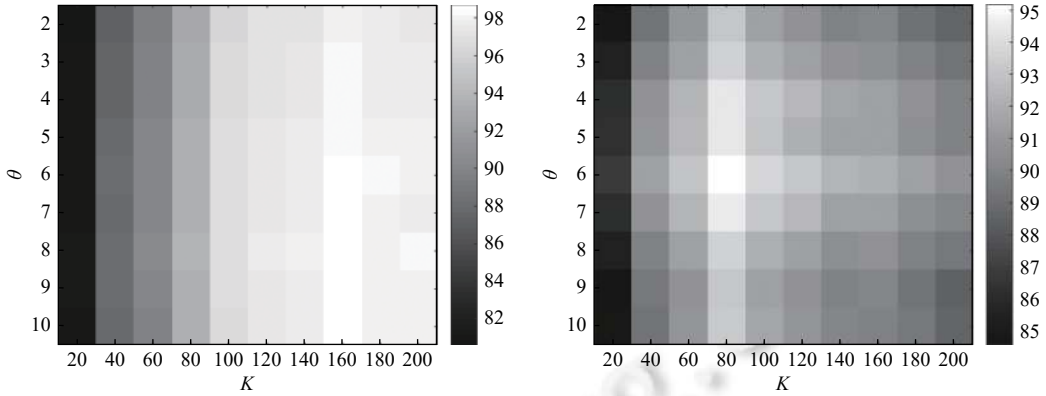


图 6 生成器空间规模和低概率阈值对行为识别正确率的影响

表 4 Grenander 测度对行为识别正确率的影响 (%)

Grenander测度	UCF101	Olympic Sports
3D CNN-LSTM	92.33	89.52
Grenander语义测度	92.50	88.85
Grenander特征测度	93.91	89.88
Grenander融合测度	**98.74**	**95.19**

(4) 时间模式的影响. 表 5 给出了不同时间模式下的消融实验, 其他参数设置为图 6 中最高正确率的条件. 可以看出: ① 在初始的时间模式, 都比基准模型的正确率高, 这可能是显式时间模式, 能够避免时间节点的歧义信息, 避免干扰在前向传递中进一步扩大. ② 同时, 可以看出慢行为的冗余去除, 异常节点的去除对于学习时间模式都是有效的, 这说明真实视频中的干扰是存在的. ③ Grenander 推理相对于去除异常后的时间模式, 在 UCF101 和 Olympic Sports 上都有明显的提高, 说明语义约束, 对于去除歧义标记的有效性.

表 5 时间模式对行为识别正确率的影响 (%)

时间模式	UCF101	Olympic Sports
3D CNN-LSTM	92.33	89.52
初始时间模式	94.30	90.68
去除冗余的时间模式	96.78	91.51
去除异常的时间模式	97.56	92.46
Grenander推理时间模式	**98.74**	**95.19**

4.6 可视化分析

图 7 给出了行为的时间模式演化过程, 包括, UCF101 数据集的 (a) 标枪投掷, (b) 篮球扣篮, (c) 骑马, 和 Olympic Sports 数据集的 (d) 篮球, (e) 撑竿跳高, (f) 挺举. 我们注意到时间模式过程中, 确实抑制了干扰的时间节点和标记. (1) 在 (b) 篮球扣篮和 (d) 篮球中, 可以通过低概率阈值去除运动中的停顿. (2) 另一种低概率的情况是摄像机运动造成的背景运动, 例如 (a) 标枪投掷和 (e) 撑竿跳高的中的节点. (3) Grenander 推理能消除一些歧义的标记, 例如在 (d) 篮球中运球和步行, 在 (f) 挺举中蹲和弯曲, 在 (a) 标枪投掷中以不同的速度地奔跑, 在 (c) 骑马中以不同的速度地骑马. 可视化结果说明, 本文方法能够鲁棒的处理行为时间模式中的各种干扰. 在 UCF101 数据集中, 采用时间状态数 K=160, 图 7(b) 篮球扣篮有 6 个时间节点, 该视频本文增量推理时间为 46 ms, 推理时间小于深度学习特征提取的时间, 这说明本文推理方法的时间是可以接受的.


初始时间模式
49 49 49 134 134 109 109 23 109 109 146 146 146 146 109 109
去除冗余的时间模式
49 134 109 23 109 146 109
去除异常的时间模式
49 109 146 109
Grenander 推理时间模式
146 109 146 109

(a) 标枪投掷 Javelin throw


初始时间模式
126 82 82 82 6 52 52 17 148 52 52 52 148 148 148 148
去除冗余的时间模式
126 82 6 52 17 148 52 148
去除异常的时间模式
126 82 52 148 52 148
Grenander 推理时间模式
148 82 52 148 52 148

(b) 篮球扣篮 Basketball dunk


初始时间模式
101 101 106 115 115 115 100 100 100 100 115 115 115 128 115 115
去除冗余的时间模式
101 106 115 100 115 128 115
去除异常的时间模式
101 115 100 115
Grenander 推理时间模式
100 115 100 115

(c) 骑马 Horse riding


初始时间模式
23 13 13 27 27 27 27 27 30 13 13 13 13 13 13 13
去除冗余的时间模式
23 13 27 30 13
去除异常的时间模式
23 13 27 13
Grenander 推理时间模式
27 13 27 13

(d) 篮球 Basketball


初始时间模式
33 27 27 35 33 33 33 33 4 4 37 4 4 4 4 4
去除冗余的时间模式
33 27 35 33 4 37 4
去除异常的时间模式
33 27 33 4
Grenander 推理时间模式
33 4 33 4

(e) 撑杆跳高 Pole vault

图 7 本文方法的时间模式演化过程

初始时间模式

26	14	40	14	40	40	8	8	8	8	8	8	8	8	8	8

去除冗余的时间模式

26	14	40	14	40	8

去除异常的时间模式

26	14	40	14	40	8

Grenander 推理时间模式

40	14	40	14	40	8

(f) 挺举 Clean and jerk

图 7　本文方法的时间模式演化过程 (续)

5 总　结

本文首次对时间图模型使用 Grenander 模式理论进行图结构优化, 提出一种 Grenander 推理优化下时间图模型 (TGM-GI). 具体来说, 本文将其中的时间图结构优化问题, 定义为在 3D CNN-LSTM 深度模型上的时间图模式推理. 首先, 采用 K-means 方法初始化 Grenander 生成器空间, 随后, 设计了相似性合并模块和低概率特征抑制模块, 抑制时序图模型中的同质冗余和异常干扰, 最后, 构建了一种时序增量形式的 Viterbi 算法, 使用融合特征约束和语义约束的 Grenander 测度, 来解决二义性视觉干扰问题. 在 UCF101 和 Olympic Sports 两个公认数据集上, 与现有多种基于深度学习的行为识别方法进行比较, 本文方法获得了最好的行为识别正确率. 本文方法优于基准的 CNN-LSTM 方法, 在 UCF101 数据集上提高 6.41%, 在 Olympic Sports 数据集上提高 5.67%. 为了全面分析本文方法的贡献, 我们进一步设计消融实验来分析, 聚类数, 低概率阈值, Grenander 测度, 时间模式的类型的影响, 并给出时间模式的可视化过程, 充分说明了 Grenander 优化的时间图模式的有效性.

References:

[1] Deng SZ, Wang BT, Yang CG, Wang GR. Convolutional neural networks for human activity recognition using multi-location wearable sensors. Ruan Jian Xue Bao/Journal of Software, 2019, 30(3): 718−737 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5685.htm [doi: 10.13328/j.cnki.jos.005685]

[2] Tang C, Wang WJ, Li W, Li GB, Cao F. Multi-Learner co-training model for human action recognition. Ruan Jian Xue Bao/Journal of Software, 2015, 26(11): 2939−2950 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4899.htm [doi: 10.13328/j.cnki.jos.004899]

[3] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4724–4733. [doi: 10.1109/CVPR.2017.502]

[4] Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6450–6459. [doi: 10.1109/CVPR.2018.00675]

[5] Li ZY, Gavrilyuk K, Gavves E, Jain M, Snoek CGM. VideoLSTM convolves, attends and flows for action recognition. Computer Vision and Image Understanding, 2018, 166: 41–50. [doi: 10.1016/j.cviu.2017.10.011]

[6] Tran D, Wang H, Feiszli M, Torresani L. Video classification with channel-separated convolutional networks. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 5551–5560. [doi: 10.1109/ICCV.2019.00565]

[7] Wang LM, Xiong YJ, Wang Z, Qiao Y, Lin DH, Tang XO, Van Gool L. Temporal segment networks for action recognition in videos. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740–2755. [doi: 10.1109/TPAMI.2018.2868668]

[8] Kukleva A, Kuehne H, Sener F, Gall J. Unsupervised learning of action classes with continuous temporal embedding. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12058–12066. [doi: 10.1109/CVPR.2019.01234]

[9] De Souza FDM, Sarkar S, Srivastava A, Su JY. Spatially coherent interpretations of videos using pattern theory. Int'l Journal of

Computer Vision, 2017, 121(1): 5–25. [doi: 10.1007/s11263-016-0913-6]

[10] Wang H, Schmid C. Action recognition with improved trajectories. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 3551–3558. [doi: 10.1109/ICCV.2013.441]

[11] Ni BB, Moulin P, Yang XK, Yan SC. Motion Part Regularization: Improving action recognition via trajectory group selection. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3698–3706. [doi: 10.1109/CVPR.2015.7298993]

[12] Jones S, Shao L. A multigraph representation for improved unsupervised/semi-supervised learning of human actions. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 820–826. [doi: 10.1109/CVPR.2014.110]

[13] Lan ZZ, Lin M, Li XC, Hauptmann AG, Raj B. Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 204–212. [doi: 10.1109/CVPR.2015.7298616]

[14] Wang LM, Li W, Li W, Van Gool L. Appearance-and-relation networks for video classification. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1430–1439. [doi: 10.1109/cvpr.2018.00155]

[15] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proc. of the Neural Information Processing Systems. Montreal: NIPS, 2014. 568–576.

[16] Zhu Y, Lan ZZ, Newsam S, Hauptmann A. Hidden two-stream convolutional networks for action recognition. In: Proc. of the 14th Asian Conf. on Computer Vision. Perth: Springer, 2018. 363–378. [doi: 10.1007/978-3-030-20893-6_23]

[17] Ji SW, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221–231. [doi: 10.1109/TPAMI.2012.59]

[18] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 4489–4497. [doi: 10.1109/ICCV.2015.510]

[19] Xiao JH, Cui XH, Li F. Human action recognition based on convolutional neural network and spatial pyramid representation. Journal of Visual Communication and Image Representation, 2020, 71: 102722. [doi: 10.1016/j.jvcir.2019.102722]

[20] Tran D, Ray J, Shou Z, Chang SF, Paluri M. ConvNet architecture search for spatiotemporal feature learning. arXiv: 1708.05038, 2017.

[21] Wang XL, Girshick R, Gupta A, He KM. Non-local neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803. [doi: 10.1109/CVPR.2018.00813]

[22] Xie SN, Sun C, Huang J, Tu ZW, Murphy K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 318–335. [doi: 10.1007/978-3-030-01267-0_19]

[23] Diba A, Pazandeh AM, Van Gool L. Efficient two-stream motion and appearance 3D CNNs for video classification. arXiv: 1608.08851, 2016.

[24] Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7082–7092. [doi: 10.1109/iccv.2019.00718]

[25] Zolfaghari M, Singh K, Brox T. ECO: Efficient convolutional network for online video understanding. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 713–730. [doi: 10.1007/978-3-030-01216-8_43]

[26] Gan C, Wang NY, Yang Y, Yeung DY, Hauptmann AG. DevNet: A deep event network for multimedia event detection and evidence recounting. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 2568–2577. [doi: 10.1109/CVPR.2015.7298872]

[27] Wang LM, Xiong YJ, Lin DH, Van Gool L. UntrimmedNets for weakly supervised action recognition and detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6402–6411. [doi: 10.1109/CVPR.2017.678]

[28] Long X, Gan C, De Melo G, Wu JJ, Liu X, Wen SL. Attention clusters: Purely attention based local feature integration for video classification. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7834–7843. [doi: 10.1109/CVPR.2018.00817]

[29] Zhou BL, Andonian A, Oliva A, Torralba A. Temporal relational reasoning in videos. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 831–846. [doi: 10.1007/978-3-030-01246-5_49]

[30] Xu DJ, Xiao J, Zhao Z, Shao J, Xie D, Zhuang YT. Self-supervised spatiotemporal learning via video clip order prediction. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10326–10335. [doi: 10.1109/CVPR.2019.01058]

[31] Wu CY, Feichtenhofer C, Fan HQ, He KM, Krähenbühl P, Girshick R. Long-term feature banks for detailed video understanding. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 284–293. [doi: 10.1109/CVPR.2019.00037]

[32] Dwibedi D, Aytar Y, Tompson J, Sermanet P, Zisserman A. Temporal cycle-consistency learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1801–1810. [doi: 10.1109/CVPR.2019.00190]

[33] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, Saenko K. Long-term recurrent convolutional networks for visual recognition and description. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 2625–2634. [doi: 10.1109/CVPR.2015.7298878]

[34] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[35] Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: Deep networks for video classification. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4694–4702. [doi: 10.1109/CVPR.2015.7299101]

[36] Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW. Action recognition in video sequences using deep Bi-directional LSTM with CNN features. IEEE Access, 2018, 6: 1155–1166. [doi: 10.1109/ACCESS.2017.2778011]

[37] Ouyang X, Xu SJ, Zhang CY, Zhou P, Yang Y, Liu GH, Li XL. A 3D-CNN and LSTM based multi-task learning architecture for action recognition. IEEE Access, 2019, 7: 40757–40770. [doi: 10.1109/ACCESS.2019.2906654]

[38] Song LF, Weng LG, Wang LF, Min X, Pan CH. Two-stream designed 2D/3D residual networks with Lstms for action recognition in videos. In: Proc. of the 25th IEEE Int'l Conf. on Image Processing. Athens: IEEE, 2018. 808–812. [doi: 10.1109/ICIP.2018.8451662]

[39] Gaidon A, Harchaoui Z, Schmid C. Actom sequence models for efficient action detection. In: Proc. of the CVPR 2011. Colorado Springs: IEEE, 2011. 3201–3208. [doi: 10.1109/CVPR.2011.5995646]

[40] Lan T, Zhu YK, Zamir AR, Savarese S. Action recognition by hierarchical mid-level action elements. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 4552–4560. [doi: 10.1109/ICCV.2015.517]

[41] Tang K, Li FF, Koller D. Learning latent temporal structure for complex event detection. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 1250–1257. [doi: 10.1109/CVPR.2012.6247808]

[42] Kuehne H, Gall J, Serre T. An end-to-end generative framework for video segmentation and recognition. In: Proc. of the 2016 IEEE Winter Conf. on Applications of Computer Vision. Lake Placi: IEEE, 2016. 1–8. [doi: 10.1109/WACV.2016.7477701]

[43] Li WX, Vasconcelos N. Complex activity recognition via attribute dynamics. Int'l Journal of Computer Vision, 2017, 122(2): 334–370. [doi: 10.1007/s11263-016-0918-1]

[44] Müller M. Information Retrieval for Music and Motion. Berlin: Springer, 2007. I-XV, 1–313.

[45] Pérez JS, Meinhardt-Llopis E, Facciolo G. TV-L1 optical flow estimation. Image Processing on Line, 2013, 3: 137–150. [doi: 10.5201/ipol.2013.26]

附中文参考文献:

[1] 邓诗卓, 王波涛, 杨传贵, 王国仁. CNN多位置穿戴式传感器人体活动识别. 软件学报, 2019, 30(3): 718–737. http://www.jos.org.cn/1000-9825/5685.htm [doi: 10.13328/j.cnki.jos.005685]

[2] 唐超, 王文剑, 李伟, 李国斌, 曹峰. 基于多学习器协同训练模型的人体行为识别方法. 软件学报, 2015, 26(11): 2939–2950. http://www.jos.org.cn/1000-9825/4899.htm [doi: 10.13328/j.cnki.jos.004899]

吴克伟(1984－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为计算机视觉, 模式识别, 人工智能.

谢昭(1980－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为计算机视觉, 模式识别, 人工智能.

高涛(1996－), 男, 硕士, 主要研究领域为计算机视觉, 模式识别, 人工智能.

郭文斌(1998－), 男, 博士, 副研究员, CCF 会员, 主要研究领域为计算机视觉, 模式识别, 人工智能.