基于三维卷积的视频语义表征与视频片段检索实验报告

GOODLAB 实验考核项目—视频检索任务

Napreth
南昌大学
dev@napreth.com

***Abstract*—本文围绕 GOODLAB 考核任务“视频语义表征与相似片段检索”展开，基于自定义三维卷积核构建轻量级视频语义特征提取框架 VideoSemanticRepresentation，实现了视频片段的语义向量化与欧氏距离检索。实验以《Bad Apple!!》视频为数据集，通过 GPU 加速实现帧级卷积与缓存复用，并在多组查询样本上验证了检索准确性与系统性能。**

***Index Terms*—视频语义表征, 三维卷积, 特征提取, 相似性检索, GPU 加速**

I. 引言

近年来，视频语义检索在智能监控、内容理解与多模态分析领域中具有重要意义。传统的基于帧差或关键帧匹配的方法难以捕获时序特征，因此需要在时间维度上进行特征建模。

本实验旨在实现一个可运行的轻量级视频语义表征系统，从原始视频中提取三维卷积特征向量，用于语义层面的片段匹配与检索。

II. 相关研究与方案分析

视频语义表征的发展大致可以分为三个层次：从低层次的物理信号特征（如光流与边缘），到中层次的运动模式与时空特征，再到高层次的语义或多模态理解。这一层次化结构决定了不同方法在表示能力、计算复杂度与可解释性上的差异。

A. 低层次：基于像素与光流的表征

早期的视频语义分析依赖于直接的视觉线索，如亮度变化、帧间差分与光流场等。典型方法包括 **Horn - Schunck** 与 **Lucas - Kanade** 光流算法，通过计算相邻帧间像素的运动向量场，估计局部速度分布。这类方法计算效率高、对噪声敏感，主要用于运动检测或目标跟踪任务。然而，低层表征仅能捕获瞬时的像素级变化，无法反映较长时间尺度上的语义关联，也不具备空间结构的表达能力。

B. 中层次：基于运动特征的统计建模

随着序列建模思想的引入，研究者开始将视频视为时间序列信号，通过统计或概率模型描述运动的动态模式。典型代表是 隐马尔可夫模型（**HMM**）、动态纹理模型（**Dynamic Texture**）等，它们通过状态转移和观测概率来刻画连续帧之间的时序依赖关系。这一层次的表征能够捕捉动作的周期性与节奏特征，但模型容量有限，难以处理复杂场景和非线性运动。

C. 高层次：基于深度学习的语义表征

深度学习的兴起使视频语义表征进入高层阶段。以卷积神经网络（CNN）为基础的模型能够同时学习空间与时间结构，形成强语义的分布式表示：

- **2D CNN**：通过对关键帧逐帧提取图像特征（如 ResNet、VGG），再在时间维度上聚合。这种方法结构简单，但忽略了帧间连续性。
- **3D CNN [1]**：在卷积核中加入时间维度（如 C3D、I3D、R3D 等），直接建模时空体积信息。以 **I3D**（**Inflated 3D ConvNet**）为代表的模型将二维卷积扩展为三维卷积，使网络同时学习空间纹理与运动动态，是当前动作识别与视频理解的主流方案。
- **Transformer** 模型：以 **TimeSformer [2]**、**ViViT** 为代表，基于自注意力机制全局建模帧间依赖，可捕捉跨时间的长程语义关系，但其计算代价极高，对显存和算力要求较大。

D. 方案选择

综合对比后，本实验选择基于自定义卷积核的三维卷积（**3D CNN**）方案。该方法兼顾了传统光流的局部可解释性与深度模型的空间 - 时间建模能力，能够在较低计算成本下实现有效的语义特征提取。通过设计针对运动方向、形状变化、反相差异和边缘信息的卷积核，本系统实现了

对视频序列时空特征的紧凑表征，并以欧氏距离作为特征空间的相似性度量，用于视频片段级的检索任务。

III. 方法设计

A. 原理与思路

视频语义表征的核心在于从连续帧序列中提取能够反映语义变化的特征。在本实验中，我们采用基于三维卷积（3D Convolution）的时空建模思想。相较于传统的二维卷积仅在空间维度上滑动，三维卷积在时间轴上引入额外的维度，使得卷积核能够同时感知相邻帧之间的动态变化，从而捕捉运动趋势、形状变换与光照变化等语义信号。

从数学角度看，三维卷积可表示为：

$$Y(t,x,y) = \sum_{\tau,i,j} W(\tau,i,j) \cdot X(t-\tau, x-i, y-j)$$

其中，W 为三维卷积核，X 为输入的时序帧序列。卷积结果反映了在局部时间-空间区域内的变化模式。当卷积核具有方向性或边缘性时，输出响应便能表征不同的运动或结构特征。

基于此思路，本实验设计了若干自定义卷积核，模拟视频中不同类型的变化：

- 运动核：检测帧间的方向性运动（左移、右移、上移、下移）；
- 形状核：利用 Laplacian 算子提取局部亮度结构变化；
- 反相核：模拟帧间全局强度翻转，强调背景与前景反差；
- 边缘核：基于 Sobel 模板检测新出现的边缘。

每一组卷积核在时间-空间上响应不同类型的动态，形成一个 7 维的特征空间。这些特征经时序聚合后可被视为视频的“语义指纹”，能够反映全局运动趋势与局部结构变化。

在检索阶段，我们将参考视频与查询片段分别转换为特征矩阵，通过滑动窗口计算欧氏距离，找到最小距离对应的区间作为匹配结果。该方案在保证可解释性和计算可控性的同时，避免了大型深度网络的训练与依赖。

B. 项目实现

根据上述原理，本项目实现了一个轻量级的视频语义表征与检索框架，命名为 **VideoSemanticRepresentation**。整体流程包括视频预处理、三维卷积特征提取、特征缓存管理与相似性检索四个阶段。

项目结构如下：

- **video.py**：负责视频读取与帧序列生成。利用 OpenCV 解码视频帧，并以 CuPy 数组的形式传输至 GPU，完成灰度化与按时间分块；
- **feature.py**：实现自定义 3D 卷积核的构造与卷积计算。通过 `cupyx.scipy.ndimage.convolve` 在 GPU 上执行三维卷积，对每个块输出 7 维特征向量，并加入哈希缓存机制避免重复计算；
- **search.py**：实现特征空间检索。使用滑动窗口计算查询片段与参考视频在特征空间的欧氏距离，输出最相似片段的时间区间；
- **___main___.py**：提供命令行接口，支持两种模式：
  - `feature` 模式：提取并保存视频特征；
  - `search` 模式（默认）：执行参考视频与查询片段的检索。

系统核心流程如下：

1) 将视频按时长（本实验设为 0.5 秒）分块；
2) 对每个分块执行多核 3D 卷积，提取运动与结构特征；
3) 将卷积响应按时间归一化并聚合为特征矩阵；
4) 对比参考视频与查询片段的特征矩阵，计算欧氏距离最小的时间窗口；
5) 输出匹配区间与距离得分。

在实际运行中，系统可自动检测缓存并复用已提取的特征，大幅缩短重复检索的计算时间。GPU 加速的三维卷积部分在 RTX 5060 上可实时处理 4K@60FPS 视频，展现出良好的性能与扩展潜力。

IV. 实验设置

A. 实验环境

- CPU: Intel Core Ultra 9 275 HX
- GPU: NVIDIA GeForce RTX 5060 Laptop GPU
- 软件: Python 3.13.8, CuPy, OpenCV, NumPy
- 系统: Windows 11 25H2

B. 数据集

使用《Bad Apple!!》作为测试视频，将原视频切分为每段 5 秒的子视频作为检索查询集。

V. 实验结果与分析

在多组查询片段中，系统成功定位出匹配区间，结果如表 I 所示。

表 I
实际检索结果（Bad Apple!! 4K60 数据集）

查询片段	匹配区间 (秒)	欧氏距离得分
s002.mp4	10.00–15.00	4.59×10^5
s015.mp4	75.00–80.00	2.73×10^5

系统能够稳定找到语义相似片段，证明三维卷积在时序特征捕捉中的有效性。

VI. 问题与改进方向

- 问题：等待 CPU 读取视频的过程中，显卡利用率偏低，算力未充分发挥；改进方向：采用双缓存机制，并行处理 CPU 视频读取与显卡的特征计算任务，提升显卡利用率。
- 问题：当前系统仅支持灰度视频的处理，无法适配彩色视频等更多场景需求；改进方向：将卷积数组维度提升一个维度，扩展对彩色视频通道（如 RGB 三通道）的兼容能力。
- 问题：自定义卷积核的适配性有限，难以匹配复杂视频的时空特征提取需求；改进方向：引入 C3D [1]、I3D 等成熟的视频专用卷积模型，通过预训练或微调的方式获取适配复杂场景的卷积核。

VII. 结论

本实验构建了一个基于三维卷积的视频语义表征系统，实现了片段级语义检索。实验验证了该方法在性能与准确性上的平衡效果，具有较高的扩展性与可解释性。

附录与项目链接

代码仓库：https://github.com/Napreth/VideoSemanticRepresentation

实验数据：https://dav.napreth.com/index.php/s/BZpkp9487w4tWwQ

参考文献

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *"Learning Spatiotemporal Features with 3D Convolutional Networks,"* in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497, doi:10.1109/ICCV.2015.510.

[2] G. Bertasius, H. Wang, and L. Torresani, *"Is Space-Time Attention All You Need for Video Understanding?"* in *Proc. 38th Int. Conf. on Machine Learning (ICML)*, vol. 139, pp. 813–824, Jul. 2021. Available: https://proceedings.mlr.press/v139/bertasius21a.html.