 DOI: https://doi.org/10.12345/bdai.v7i3.38187

Temporal Feature Alignment and Semantic Association Algorithm for Video-Text Cross-modal Retrieval

Binbin Zhang Sitang Ren

Information Engineering University of Cyberspace Force, Zhengzhou, Henan, 450001, China

Abstract

Video-text cross-modal retrieval is an important task in multi-modal understanding. Existing methods generally have problems such as temporal feature misalignment and inaccurate semantic association. To address these deficiencies, a multi-scale temporal feature alignment and hierarchical semantic association modeling method is constructed. Through temporal segment extraction, dynamic offset correction, shared semantic space construction, fine-grained relationship perception, and temporal consistency fusion, the modality matching is optimized. Experimental results show that this method effectively reduces temporal alignment errors and improves retrieval recall rates, demonstrating significant advantages in both temporal alignment accuracy and semantic association accuracy.

Keywords

video-text retrieval; temporal feature alignment; semantic association; c-modal representation; feature fusion

视频 – 文本跨模态检索中的时序特征对齐与语义关联算法

张彬彬 任思堂

网络空间部队信息工程大学，中国 · 河南 郑州 450001

摘 要

视频-文本跨模态检索是多模态理解的重要任务，现有方法普遍存在时序特征错位、语义关联不精准等问题。针对上述缺陷，构建多尺度时序特征对齐与分层语义关联建模方法，通过时序片段提取、动态偏移校正、共享语义空间构建、细粒度关系感知与时序一致性融合实现模态匹配优化。实验结果表明，该方法有效降低时序对齐误差，提升检索召回率，在时序对齐精度与语义关联准确性上均具备明显优势。

关键词

视频-文本检索；时序特征对齐；语义关联；跨模态表征；特征融合

1 引言

视频与文本异构数据的跨模态检索在内容理解领域应用广泛，传统方法多聚焦全局特征匹配，忽略时序结构差异与细粒度语义关联，易出现时序错位、语义匹配偏差等问题。为提升检索精度与时序一致性，围绕时序特征对齐与语义关联展开研究，构建多尺度时序建模与分层语义关联算法，实现视频时序信息与文本语义序列的匹配贴合，为高效可靠的跨模态检索提供技术支撑。

2 视频文本跨模态检索与时序语义关联技术概述

视频 - 文本跨模态检索以异构模态数据的时序建模与语义关联为核心，依托深度表征学习搭建视频视觉时序序列与文本语义序列的统一映射空间 [1]。视频端借助多尺度时序编码捕获帧间运动依赖、片段级时序演变与整体时序布局，文本端依靠序列建模完成词汇、短语及语句层次的语义拆解，生成模态内部规整的特征表述。当前技术以跨模态对比学习为基础架构，围绕时序位置匹配与语义概念关联两类核心目标，借助模态共享投影与时序一致性约束缩减异构特征分布差距，依靠细粒度语义交互发掘实体、动作与场景的跨模态对应联系。时序特征对齐打破静态特征匹配限制，把视频动态变化与文本时序描述紧密结合，语义关联打通模态间语义隔阂，完成从粗粒度整体匹配到细粒度单元关联的过渡，为复杂视频内容的文本检索提供技术支持，带动跨模态理解朝时序感知与语义深度融合方向发展。

3 视频 – 文本跨模态检索存在时序特征错位与语义关联不精准问题

视频 - 文本跨模态检索场景中，异构模态数据在时序维度与语义维度存在显著表征差异，直接引发时序特征错位与语义关联偏差问题 [2]。视频数据具备连续动态的时序结构，帧间变化呈现非线性演化特征，文本描述多为离散化语义符

【作者简介】张彬彬（1993–），男，中国河南新蔡人，工程师，硕士，从事智能化检索（大数据分析）研究。

号序列，二者在时序长度、时序节点对应关系上难以形成自然匹配，易出现时序偏移、时序节奏不匹配等错位现象。现有表征方式多侧重全局特征映射，缺少对局部时序片段与细粒度语义单元的关联建模，实体、动作及场景信息在跨模态传递过程中易出现语义弱化与对应偏差，难以实现稳定语义绑定。跨模态分布差异进一步加大时序对齐误差，检索模型难以兼顾时序一致性与语义相关性，直接制约跨模态检索的精度与整体鲁棒性。

4 多尺度时序特征对齐与分层语义关联建模

4.1 视频时序片段化特征提取与文本序列结构化编码

视频时序片段化特征提取以连续视觉帧序列为对象，按固定长度划分原始视频，将 1200 帧视频流分为 30 个独立时序单元，每个单元含 40 帧连续视觉信息，经三维卷积网络初步提取局部时空特征，在通道维度生成 512 维特征向量，保留帧间运动与空间布局信息。接着采用滑动窗口机制，以 20 帧步长对相邻片段重叠采样，强化时序上下文连续性，避免边界信息丢失，同时通过时序池化聚合关键特征，压缩冗余、突出时序规律。文本序列结构化编码利用预训练语言模型对输入文本逐层解析，将文本映射为 768 维语义特征向量，按语义粒度划分特征，通过位置编码注入顺序信息，构建适配视频时序结构的特征体系，使文本语义表征有可对齐的时序属性，为跨模态时序匹配奠基。

4.2 动态时序偏移校正与帧级时序特征匹配

动态时序偏移校正是基于时序注意力机制设计自适应对齐模块来估计视频片段以及文本序列之间的时间差并对其进行矫正，设置时间偏移上限为 15 帧，对于超过此范围的错位特征赋予一定的权重进行修正，使用动态规划求解最佳时序路径从而减小不同模态之间的时序差距。而这个过程中的时间偏移是可学习的并且可以随着训练不断更新其对应的权重，以此降低由于非线性的时序变化造成的对齐误差同时又不会忽略掉视频的关键帧以及文本的重要部分，提高了时间上的鲁棒性。帧级时序特征匹配是以得到的视频帧特征和文本结构化特征作为输入，在一个相同的空间下进行一一对应的操作，将帧特征维度以及文本特征维度都转换到 256 维度上，然后用余弦相似度来进行衡量它们之间的相似程度选出那些相似度比较高的特征作为一个初步的匹配集合作为下一步处理的基础。加入时序连续性约束改进匹配结果，去掉分散且不符合预期的匹配，加强帧级特征时间一致性，使视频动态时序信息与文本语义顺序更好对应，解决时序错位问题。

4.3 跨模态共享语义空间构建与特征投影对齐

跨模态共享语义空间构建以统一异构模态特征分布为目标，通过多层非线性映射网络将视频时序特征与文本语义特征投射至同一度量空间。先构建维度为 512 的共享特征子空间，分别设计视频分支全连接投影层与文本分支全连接投影层，两层网络均采用三层线性变换结构，每层隐藏层维度依次设置为 768、512、256，激活函数采用 GELU 完成非线性特征转换。引入特征分布对齐损失约束模态间分布差异，构建损失函数：

$$L_{proj}=\|M_vF_v-M_tF_t\|_2^2 \quad (1)$$

其中 M_v 为视频投影矩阵，M_t 为文本投影矩阵，F_v 为视频时序片段特征，F_t 为文本序列编码特征。在训练过程中固定批处理尺度为 32，每轮迭代更新一次矩阵参数，通过梯度下降逐步缩小特征分布距离。

4.4 实体动作关系感知的细粒度语义关联计算

实体动作关系感知的细粒度语义关联计算聚焦视频视觉单元与文本语义单元的精细化匹配，突破全局特征匹配的局限性。先对视频特征进行区域拆分，将单帧特征划分为 36 个局部视觉块，提取目标实体与运动轨迹特征，再对文本特征执行句法解析，分离名词性实体特征与动词性动作特征，形成多组细粒度特征对。构建关联度量函数：

$$S_{sem}=\sum(f_{ent}\cdot f_{act})+\alpha\cdot\|f_{ent}-f_{act}\|_1 \quad (2)$$

其中 f_{ent} 表示实体特征向量，维度为 128，f_{act} 表示动作特征向量，维度同样为 128，α 为关系加权系数，取值固定为 1.2。计算过程中逐一遍历所有实体与动作组合，保留关联得分高于阈值 6 的有效匹配对，过滤低相关噪声组合。通过关系感知注意力机制对有效对进行加权融合，突出强关联语义单元的贡献，同时记录各匹配对对应的时序位置信息，使语义关联结果与时序结构保持绑定，实现从单一特征匹配向结构化关系关联的升级，提升跨模态语义对应的完整性与精准度。

4.5 时序一致性约束下的跨模态相似度融合

时序一致性约束下的跨模态相似度融合综合时序对齐结果与语义关联得分，构建兼顾时序合理性与语义相关性的统一度量机制。先将帧级时序匹配得分与细粒度语义关联得分分别归一化至固定区间 0 到 10，再通过融合函数：

$$S_{final}=\beta S_{temp}+\gamma S_{sem} \quad (3)$$

完成数值融合，其中 S_{temp} 为时序对齐得分，S_{sem} 为语义关联得分，β 与时序权重相关取值 2.5，γ 为语义权重取值 3.8。引入时序平滑惩罚项抑制突变匹配结果，惩罚项计算基于相邻帧位置差值，当单步位置偏移超过 12 帧时自动提升惩罚系数，削弱异常匹配对整体得分的干扰 [4]。融合过程以视频片段为基本单元，对长度为 40 帧的片段执行滑动计算，步长设置为 10 帧，逐段输出融合相似度。最终将所有片段得分进行累加聚合，形成视频与文本间的整体相似度数值，该数值直接作为跨模态检索排序依据，在保证语义匹配精度的同时强化时序结构的连续性，有效改善时序错位导致的检索排序混乱问题。

5 实验设计与结果分析

5.1 实验数据集与评价指标设定

实验选用面向视频文本跨模态检索的公开数据集开展验证工作，数据集包含经过时序标注的视频片段与文本描述对，总视频数量为 2392 条，每条视频平均时长 450 帧，配套文本描述共计 14352 条，每条文本包含 28 至 62 个语义单元，覆盖场景、实体、动作与时序变化等多维度信息。数据集按固定比例划分为训练集、验证集与测试集，样本数量依次为 1914 条、239 条、239 条，保证各子集在场景复杂度与时序多样性上分布均衡。评价体系围绕检索精度与时序匹配效果构建，采用 Recall@1、Recall@5、Recall@10 作为核心检索评价指标，分别记录排序首位、前五位与前十位内正确匹配样本的召回情况，同时引入时序对齐误差指标，以帧为单位计算视频与文本对应位置的平均偏移量。所有指标均在单一检索任务上独立计算，测试阶段输入固定维度的特征向量，批量处理规模设为 32，保证指标计算过程的一致性与稳定性，全面反映模型在时序对齐与语义关联层面的综合性能。

5.2 对比方法与实现细节说明

实验选取多种主流跨模态检索模型作为对比基准，涵盖基于全局特征映射、时序注意力机制以及语义图结构建模的经典方法，所有模型均在统一硬件环境与软件框架下完成复现与调试。硬件平台采用单卡显存为 24GB 的加速计算卡，内存容量设为 128GB，训练批次大小固定为 16，迭代总轮次设置为 80 轮，初始学习率设为 0.0001，每 20 轮执行一次学习率衰减操作。模型输入层面统一规范视频特征长度与文本编码维度，视频片段采样长度为 40 帧，文本编码长度统一填充至 80 个语义单元，避免输入尺度差异对实验结果造成干扰 [5]。实现过程中对各对比模型的投影层维度、注意力头数、时序窗口长度等关键参数进行标准化配置，投影层维度统一设为 512，注意力头数固定为 8，时序窗口长度设为 20 帧。训练阶段采用相同的数据增强策略与优化器类型，测试阶段保持检索排序规则一致，通过控制变量的方式排除外部条件干扰，确保对比结果能够客观反映时序特征对齐与语义关联算法的有效性。

5.3 时序对齐精度与检索性能结果分析

时序对齐精度测试以帧级偏移量为核心衡量依据，对测试集 239 条样本逐一对齐结果进行量化统计，实验测得平均时序偏移量为 8.6 帧，相较于对比模型平均 15.3 帧的偏移量大幅降低，其中偏移量≤ 10 帧的样本占比显著提升，有效验证时序偏移校正模块的有效性。检索性能测试围绕核心评价指标展开，在测试集上测得 Recall@1 数值为 42.7，Recall@5 数值为 68.3，Recall@10 数值为 79.5，各指标均优于对比模型对应数值，其中 Recall@1 较最优对比模型提升 7.2，Recall@10 提升 6.8。测试过程中按视频时序复杂度分层统计，对时长≥ 600 帧的长时序视频，平均时序偏移量控制在 10.2 帧，Recall@10 数值为 75.3，对时长＜ 300 帧的短时序视频，平均偏移量降至 6.9 帧，Recall@10 数值达 84.1，清晰呈现模型在不同时序尺度视频检索中的适配性，同时验证多尺度时序对齐与分层语义关联机制能够有效提升检索精度与时序匹配稳定性。

5.4 消融实验与模块有效性验证

消融实验采用逐步移除核心模块的方式，依次验证各模块对模型性能的贡献，实验基于相同硬件环境与测试集，保持其他参数不变，逐一对各模块进行单独消融测试并记录性能变化。移除动态时序偏移校正模块后，平均时序偏移量升至 16.8 帧，Recall@1 降至 32.1，Recall@10 降至 67.3，表明该模块可有效缓解时序错位问题；移除细粒度语义关联计算模块后，Recall@1 降至 35.8，Recall@10 降至 70.6，语义匹配误差增加 12.4，验证细粒度关联对语义精准度的提升作用；移除时序一致性融合模块后，检索排序混乱度提升，Recall@5 降至 58.9，平均偏移量波动增至 21.3 帧。实验同时对各模块组合效果进行测试，完整模型较单一模块性能最优的组合提升 8.7，证明各模块协同作用能够实现时序对齐与语义关联的双重优化，进一步验证所提算法各核心模块的必要性与有效性。

6 结语

视频 - 文本跨模态检索中的时序与语义问题展开研究，提出多尺度时序对齐与分层语义关联建模方案，从特征提取、时序校正、空间构建、语义计算到相似度融合形成完整体系。实验验证各模块有效提升时序对齐精度与检索性能，改善异构模态匹配效果。该方法强化了时序感知与语义深度融合，可为视频内容检索、多模态数据管理等场景提供可行技术参考。

参考文献

[1] 董闯,栗伟,巴聪,等. 基于联合嵌入空间的视频文本检索研究综述[J].中国图象图形学报,2025,30(05):1220-1237.

[2] 王盛,宋向辉,胡世雄,等. 一种基于交叉注意力机制的跨模态视频-文本检索模型[J].安全、健康和环境,2025,25(03):20-26.

[3] 习怡萌,刘立波,邓箴,等. 多级跨模态对齐的文本检索视频方法研究[J].中文信息学报,2025,39(02):111-122.

[4] 习怡萌,邓箴,刘倩,等. 跨模态信息融合的视频-文本检索[J].计算机应用,2025,45(08):2448-2456.

[5] 涂荣成,毛先领,孔伟杰,等. 基于CLIP生成多事件表示的视频文本检索方法[J].计算机研究与发展,2023,60(09):2169-2179.