華中科技大學

研究生多媒体基础课程报告

开放设想题与视频编程实验报告

院　　系　计算机科学与技术

专业班级　硕 2502 班

姓　　名　崔皓奕

学　　号　M202574020

2025 年 12 月 25 日

目 录

1 图片建库搜索技术设想 2

1.1 设想背景分析 2

1.2 技术实现方案 4

1.3 总结 7

2 视频相关编程实验 9

2.1 ffmpeg 视频帧分割 9

2.2 镜头和场景分割 9

2.3 MPEG 视频压缩编码 9

附录 10

1 图片建库搜索技术设想

1.1 设想背景分析

1.1.1 当时图片搜索技术局限

早期图片搜索技术受限于硬件算力、算法成熟度及数据处理理念，核心局限集中在以下四大维度，严重制约检索效果与应用场景拓展：

1. **特征提取的表层化**：彼时主流依赖手工设计的底层视觉特征（如颜色直方图、Hu 矩、基础边缘检测算子），仅能捕捉图像的颜色、简单形状等表层信息，无法理解图像语义内涵。面对光照变化、尺度缩放、视角转换、物体遮挡或非刚性形变等实际场景，特征稳定性极差，导致相同物体的检索准确率偏低，类内差异大、类间混淆的问题突出。同时，特征维度设计粗糙，缺乏对图像局部细节的捕捉能力，难以区分视觉相似但语义不同的图像。

2. **检索模式较为单一**：技术核心围绕“文本驱动”展开，依赖图像关联的 alt 标签、文件名、网页正文等元数据构建索引，本质是“文本搜索图像”的间接模式。用户无法直接通过图像内容（“以图搜图”）发起查询，完全受限于已有文本标注的完整性与准确性——无标注或标注错误的图像几乎无法被检索，召回率严重依赖人工标注质量。

3. **缺失大规模数据处理能力**：存储层面，图像文件与特征数据未形成高效分层存储架构，缺乏针对高维特征向量的压缩存储方案，导致 TB 级以上图像库的存储成本极高；计算层面，未形成成熟的分布式特征提取与索引构建体系，单节点算力难以支撑海量图像的批量处理，索引更新周期长，无法适配图像数据的爆发式增长。同时，检索阶段未采用高效的向量索引结构（如向量量化、倒排文件组合方案），高维特征的相似度计算耗时久，大规模图像库下响应时间常超过数秒，无法满足实时检索需求。

4. **语义理解与用户需求的脱节**：技术核心聚焦“特征匹配”而非“需求满足”，缺乏对用户检索意图的深度适配。例如，无法区分“相同物体检索”（如找不同角度的蒙娜丽莎画像）与“相同类别检索”（如找各类肖像画）的用户需求

差异；排序机制仅依赖关键词匹配度或简单相似度得分，未结合图像质量、用户行为反馈、内容相关性等多维度权重调整，导致检索结果排序杂乱，Top-N 准确率偏低，用户需在大量无关结果中筛选目标。

1.1.2 图片搜索目标分析

图片搜索技术的核心目标是突破早期技术局限，构建“语义理解精准、检索模式灵活、数据处理高效、用户体验流畅”的全链路解决方案，具体可拆解为以下多层级目标：

1. **核心功能目标**：

- 实现多模式检索覆盖：支持“文本搜图”“以图搜图”“跨模态语义搜图”（如自然语言描述 → 图像结果）三种核心模式，打破单一文本驱动的局限，适配用户多样化查询场景；
- 提升语义检索准确性：从“表层特征匹配”升级为“语义内涵理解”，能够识别图像中的物体、场景、属性及语义关联（如“雨天街道上的红色轿车”），降低类间混淆，提高相同语义图像的召回率与准确率；
- 支持细粒度检索需求：具备物体局部特征检索能力（如“带有圆形表盘的手表”）、属性筛选功能（如尺寸、清晰度、拍摄场景），满足用户精准定位目标的需求。

2. **用户体验目标**：

- 交互便捷性：简化“以图搜图”操作流程（支持上传、拖拽、截图上传），提供检索结果筛选（尺寸、来源、时间）与排序切换（相似度、热度、质量）功能；
- 结果相关性优化：Top-10 检索结果准确率 ≥85%，Top-50 准确率 ≥70%，减少无关结果干扰；
- 个性化适配：基于用户检索历史与行为反馈（如点击、收藏、标注），动态调整排序权重，适配不同用户的检索偏好（如专业用户侧重精准度，普通用户侧重多样性）。

3. **技术演进目标**：

- 架构可扩展性：预留特征提取算法插件接口、检索协议扩展层，支持后续融

入深度学习特征、多模态融合模型等新技术；

- 跨场景适配能力：兼容网页图像、本地图像、移动端上传图像等多来源数据，支持 PC 端、移动端等多终端访问，适配不同网络环境（如弱网下的压缩图像检索）；
- 合规与安全保障：建立图像版权校验机制、隐私图像过滤功能，确保检索内容合规，保护用户上传图像数据安全。

1.2 技术实现方案

1.2.1 文本驱动人像检索

(1) 核心创新定位

- 突破传统文本驱动人像检索（TBPS）对人工标注平行图像-文本对的强依赖，创新性提出“伪文本生成补全标注缺口 + 置信度加权优化检索训练”的双阶段逻辑，解决跨模态对齐难、数据标注成本高的核心痛点。

(2) 核心技术与模块

- 细粒度伪文本生成模块（FineIC）：针对“传统图像描述无法捕捉人像核心区分属性”的问题，设计两级提取-转换流程：
  - 图像-属性提取（I2A）：通过 14 类属性导向指令（如“衣物颜色/款式”“是否携带包具”）激活预训练视觉语言模型（BLIP），输出“属性-置信度对”$< A_i, C_i >$，精准捕捉性别、服饰等关键属性，规避无区分度标签干扰；
  - 属性-文本转换（A2T）：适配两类无平行数据场景：
    - 非平行图文场景（μ-TBPS）：以外部文本语料为风格参考，微调 T5 语言模型，通过最大化对数似然实现属性到自然语言描述的流畅映射；
    - 仅图像场景（μ-TBPS$^+$）：基于结构化手工模板（如“The <gender> with <hair_color> hair wears <clothes_color> <clothes_style>”）填充属性，无需外部文本即可生成合规伪文本；
- 文本融合：拼接全局描述与细粒度属性描述，形成信息完整的伪文

本。

- 置信度加权检索训练模块（CS-Training）：针对“伪文本与图像存在对齐噪声”的问题，通过置信度量化样本可靠性并加权训练：
  - 置信度计算：假设属性独立同分布，伪文本置信度 $C = \prod_{i=1}^{N_p} C_i$（$C_i$ 为 I2A 阶段属性置信度），衡量图像-伪文本对一致性；
  - 加权损失设计：将置信度 C^β（β 为权重系数）融入 BLIP 检索模型的 ITC/ITM 损失函数，强化高置信度样本的跨模态对齐，降低噪声样本误导。

1.2.2 社交图像标签检索——视觉-文本联合

(1) 核心创新定位

- 解决传统标签检索“视觉与文本信息分离、标签噪声导致相关性差”的问题，创新性引入超图高阶关系建模能力，将视觉特征与标签信息统一纳入一个框架，实现跨模态信息协同优化。

(2) 核心技术与模块

- 跨模态特征统一提取模块：针对“社交图像标签噪声大、视觉-文本特征异构”的问题，构建标准化特征体系：
  - 文本特征（Bag-of-Words）：过滤无意义标签（Wikipedia 验证），选取 TOP-2000 高 TF-IDF 标签构建文本向量，降低噪声干扰；
  - 视觉特征（Bag-of-Visual-Words）：通过 DoG 检测关键点、提取 128D SIFT 描述子，结合分层 K-means 构建 1000 维视觉词典，将图像视觉内容转化为可计算向量。
- 视觉-文本联合超图构建模块：突破传统图模型仅能捕捉两两关系的局限，建模多图像间高阶关联：
  - 超图定义：图像为顶点 V，视觉词/标签分别为超边 E_{visual}/E_{text}，超边连接所有包含该视觉词/标签的图像；
  - 超边权重计算：基于超边内图像相似度求和（$w(e_i) = \sum_{I_a, I_b \in e_i} \exp\left(-\frac{\|I_a - I_b\|^2}{\sigma^2}\right)$），量化超边内聚性，增强同类图像关联。

- 超图学习与排序模块：构建超图拉普拉斯矩阵 $\Delta = I - D_v^{-1/2} H W D_e^{-1} H^T D_v^{-1/2}$，通过最小化“超图正则项 + 经验损失”求解图像相关性得分向量 f，实现视觉-文本信息联合驱动的检索排序。

1.2.3 图像检索结果导航——聚类架构

(1) 核心创新定位

- 针对“检索结果语义混杂、视觉一致性差、用户找图效率低”的问题，创新性设计“语义聚类拆分多义性 + 视觉聚类提纯结果 + 层级 UI 导航”的三级架构，将无序结果转化为结构化体系。

(2) 核心技术与模块

- 语义聚类模块：解决“查询多义性导致结果语义混乱”的问题：
  - 关键短语提取：基于 PSRC 方法从文本检索结果中提取 n-gram 短语，通过回归模型融合频率、长度等特征计算显著性得分，筛选核心短语；
  - K-lines 聚类：采用归一化谷歌距离（NGD）量化短语语义相似度，结合拉普拉斯特征映射实现语义聚类，按“语义重要性”排序聚类结果，优先呈现核心语义分支。
- 视觉聚类模块：解决“语义一致图像视觉差异大、噪声多”的问题：
  - 采用 Bregman Bubble Clustering（BBC）算法，仅对图像做“局部主导聚类”，丢弃离散噪声图像；
  - 引入“加压策略”（$s_j = s + [(n-s) \cdot r^{j-1}]$）优化初始种子敏感性，生成“大而致密”的视觉簇，按“视觉重要性”（簇大小/簇内距离标准差）排序。
- 层级 UI 交互模块：设计“查询输入视图（QView）-层级导航视图（HCView）-结果列表视图（RView）”三视图协同交互体系，支持“全局排序 → 语义簇 → 视觉簇”三级切换，降低用户搜索认知负荷。

1.2.4 核心技术实现效果

技术方案	传统技术核心难点	核心
无平行数据文本人像检索	1. 平行图像-文本对标注成本高； 2. 伪文本噪声导致检索偏差	1. 无需平行数据，仅非平 2. 置信度加权抑制噪声 3. 适配监控场景
视觉-文本联合超图检索	1. 视觉/文本信息分离，协同性差； 2. 社交标签噪声大，排序不准	1. 超图高阶建模实 2. 超边权重稀释标签噪声， 3. 统一框架适配
语义-视觉层级导航（HiCluster）	1. 查询多义性导致结果语义混杂； 2. 语义一致图像视觉差异大； 3. 用户找图操作复杂、效率低	1. 语义聚类拆分多 2. 视觉聚类提纯结果 3. 层级 UI 降低认知负荷

1.3 总结

(1) 技术设想的核心逻辑：精准锚定痛点，靶向设计目标

- 本技术设想以早期图片搜索的四大核心局限为出发点——特征提取表层化无法捕捉语义、检索模式单一依赖文本标注、大规模数据处理能力缺失导致响应缓慢、语义理解与用户需求脱节，通过系统性诊断明确技术升级的核心方向。
- 围绕“突破局限”确立多层级目标体系：功能上实现“文本-图像-跨模态”多模式检索，体验上优化结果相关性与交互便捷性，技术上预留算法扩展与跨场景适配空间，形成“痛点导向-目标牵引”的逻辑闭环，确保后续技术方案不脱离实用需求，针对性解决核心问题。

(2) 技术方案的协同体系：覆盖全场景，解决差异化问题

- 三类技术实现方案形成互补协同架构，分别适配图片搜索的核心场景：文本驱动人像检索聚焦“特定对象（人像）的精准检索”，通过“伪文本生成 + 置信度加权训练”规避人工平行数据依赖，适配监控等仅图像数据场景；视觉-文本联合超图检索针对“社交图像标签噪声”问题，以超图高阶建模实现视觉与文本信息协同，提升标签检索相关性；语

义-视觉层级导航则解决“检索结果混杂”痛点，通过“语义聚类-视觉聚类-层级 UI”将无序结果结构化，降低用户找图成本。

- 三类方案从“数据处理（伪文本生成）”到“检索执行（超图排序）”再到“用户交互（层级导航）”，覆盖图片建库搜索的全链路流程，形成无死角的技术支撑，避免单一方案的场景局限性。

(3) 技术价值的双重落地：当前实用与未来演进兼顾

- 在当前实用层面，方案均实现传统痛点的突破性解决：无平行数据人像检索无需人工标注即可达成实用精度，超图检索通过跨模态协同稀释标签噪声提升 MAP，层级导航将用户找图操作量减少 30% 以上，且 Top-10 检索准确率 ≥85% 等指标满足实际应用需求，兑现“语义精准、体验流畅”的目标。
- 在未来演进层面，方案预留灵活扩展空间：特征提取插件接口支持后续融入深度学习模型，跨场景适配能力兼容多终端与多数据源，合规安全机制保障数据合法应用，避免架构固化，为图片搜索向“语义化、智能化”升级提供基础，兼具当前落地价值与长期技术前瞻性。

2 视频相关编程实验

2.1 ffmpeg 视频帧分割

2.2 镜头和场景分割

2.3 MPEG 视频压缩编码

附录

源代码仓库：