摘要

面对互联网上存在的大量图像数据，视觉搜索技术（基于视觉特征的图像搜索）已成为解决图像检索需求的重要技术手段，与基于文本的检索方法相比表现出更好的便捷性与客观性。随着移动互联网的发展和具有拍摄功能的手机的普及，移动视觉搜索正成为视觉搜索的一种重要形式。现有移动视觉搜索方法中通常采用单张图像作为查询进行图像检索，然而单张图像本身具有的某些缺陷性可能导致图像检索效果不佳，如单张图像的拍摄视角受限，拍摄中可能引入的图像模糊、光照不均、背景杂乱、前景遮挡等。针对上述问题，本文提出引入视频作为查询进行移动视觉搜索。一段视频能够从多个视角对查询目标物体进行视觉信息的描述，且能够通过多帧克服某单帧中可能出现的模糊等问题。

然而，移动视觉搜索条件下，引入视频作为查询面临着以下两方面的重要挑战：

首先，受限的移动无线网络带宽。在移动视觉搜索中，各种数据的传输均需经由带宽有限、带宽不稳定的移动无线网络。而一段视频片段，所对应的数据开销相对庞大，直接传输查询视频会给移动网络带来较大压力，且会产生较高的查询延迟，影响移动视觉搜索应用的用户体验。

其次，查询视频中的噪声和干扰信息影响视觉搜索效果的提升。在查询视频对目标物体进行充分的视觉信息描述的同时，其拍摄过程中会不可避免地引入诸多噪声和干扰信息。噪声和干扰信息将会在基于视觉特征的图像检索过程中造成特征误匹配，不利于查询视频对目标物体视觉信息充分性描述的优势的发挥，影响视觉搜索效果的提升。

针对以上挑战，本文提出在移动端对查询视频提取视觉特征，形成紧凑的特征描述子，以适应有限的移动网络带宽。此外，为有效提高查询视频紧凑特征表示的可区分力，本文提出对查询视频所提取的特征进行可区分力的判断，仅选取最具区分力的特征参与特征表示，抑制噪声和干扰特征的负面影响，以提升视觉搜索准确度。本文具体贡献包括从全局和局部两个角度对查询视频形成具有高区分力的紧凑特征表示：

第一，本文提出使用 Fisher 向量对视频局部特征进行聚合，形成一个具有高区分力的查询视频紧凑全局特征表示。为有效提高 Fisher 向量的可区分力，并降低聚合过程的计算复杂度，本文提出了选择性局部特征聚合方案，仅选择关键帧中的有效特征参与 Fisher 向量聚合。针对视频帧内的特征选择，本文提出了基于

视频相邻帧时域相关性与特征空域属性相结合的特征可区分力评价方法，基于特征可区分力实现特征排序及筛选。针对查询视频的关键帧选取，本文提出了基于关键帧对视频内容覆盖约束与关键帧质量相结合的关键帧提取算法，实现关键帧对视频视觉内容高效覆盖的同时，保证图像检索的高准确度。实验表明，上述特征选择与关键帧提取算法能有效提高查询视频 Fisher 向量的可区分力，且该选择性聚合的查询视频 Fisher 向量，与单张查询图像的 Fisher 向量具有同等的数据开销的前提下，带来了显著的视觉搜索效果提升，在百万规模的数据集上，基于图像查询的图像检索 mAP 为 63.2%，基于视频查询的图像检索 mAP 为 77.6%。

第二，本文提出了多关键帧协同的查询视频局部特征表示及特征匹配方法。对查询视频进行局部特征表示，能够从具体细节角度提供目标物体的局部视觉信息描述，有助于更精准的图像间特征匹配，能进一步提升图像检索准确度。本文提出同时从查询视频的多个关键帧中选取若干局部特征，实现多视角下的目标物体局部特征描述。并且，本文提出基于多关键帧协同的局部特征匹配方法，对多个关键帧与数据集图像间的特征匹配结果进行融合，提高特征匹配的准确性与鲁棒性。实验结果表明，基于多帧的局部特征匹配与仅基于单帧的特征匹配相比，具有更高的准确性。此外，基于局部特征匹配对图像检索结果进行重排序，在百万规模数据集上，能够带来 7 个百分点的 mAP 提升。

本文提出的方法实现了对查询视频高区分力且紧凑的特征表示，能够在无线网络的有限带宽下，实现高准确度的移动视觉搜索。与基于单张图像的方法相比，对应相等的数据开销，本文提出的基于视频查询的移动视觉搜索方法，在图像检索效果方面表现出明显的优越性。

关键词：查询视频，移动视觉搜索，视频特征表示，选择性聚合