密级:_____

中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

移动视频检索关键技术研究

作者姓名：刘武

指导教师：李锦涛 研究员

中国科学院计算技术研究所

学位类别：工学博士

学科专业：计算机应用技术

研究所：中国科学院计算技术研究所

2015 年 5 月

Research on Key Techniques of Mobile Video Search

By

Wu Liu

A Dissertation Submitted to

The University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Doctor of Philosophy

Institute of Computing Technology

May, 2015

声　　明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：　　日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：　　导师签名：　　日期：

摘 要

移动视频检索技术是视频检索领域中一个前沿的研究课题。近年来，移动设备的飞速发展，改变了互联网上视频内容的产生，以及人们检索和观看视频的方式。移动设备的便携性和无处不在的网络接入能力使其逐渐成为主要的视频访问和查询入口，而移动设备上丰富的传感器原件，也使得移动视频检索过程更加自然、有效。因此，移动视频检索具有巨大的市场需求和应用前景。但是，现有移动视频检索相关工作大多沿袭了传统桌面视频检索技术，忽略了移动视频检索的特有挑战和需求。不同于传统的桌面视频检索，移动视频检索主要面临如下挑战：1）查询视频受拍摄环境影响产生严重图像形变和音频噪声；2）移动设备计算性能有限；3）移动网络带宽限制；4）移动用户更重视视频检索体验。

针对移动视频检索的特点与挑战，本文对移动视频检索中的关键技术进行了研究。在移动设备上，我们主要研究了移动视频签名快速生成与加速技术；在服务端，我们研究了针对大规模视频数据进行快速检索的音/视频分层哈希索引与渐进式匹配技术。为提升用户搜索体验，我们还针对移动视频检索的结果展示研究了基于视觉-语义深度嵌入的查询相关视频缩略图选择技术。在此基础上，我们研发了一套新颖的实时移动视频检索系统。本论文的主要贡献总结如下：

1. 基于渐进式传输的视频签名生成与加速方法

移动视频检索的特有挑战要求移动视频签名具有计算简单、存储高效、区分性强、易传输和易索引等特点。针对现有视频签名计算复杂、区分能力差且传输数据量较大的问题，本文提出了一种多模态的紧凑视频签名生成方法和一种基于 Hessian 显著度加权融合的渐近式视频签名传输方法。该哈希签名方法融合了视觉哈希码和音频指纹特征，充分挖掘了不同模态特征之间的互补性，有效应对查询视频在拍摄过程中造成的图像形变和音频噪声，提高了检索精度。同时渐近式传输方法能大幅减少网络传输数据量，从而显著提高移动视频检索效率和识别比特率。经实验验证，该算法每秒只需要传输 0.88KB 的特征，与已有最好方法相比，在提高查准率的情况下，减少了 33.5%的查询延迟。

2. 音/视频分层哈希索引与渐进式视频匹配方法

针对现有二进制哈希索引方法只能处理单模态特征且检索精度不高的问题，本文提出了一种音/视频分层哈希索引与匹配方法。该方法通过音/视频分层过滤策略，高效地融合了视觉和音频两种互补的模态特征，以充分挖掘多模态特征之间关联，显著提高了检索精度。针对移动视频检索中子序列匹配问题，我们提出了利用二分图转换和最大匹配算法实现查询视频与源视频的渐进式匹配。该算法可对查询视频进行精确定位，并伴

随查询视频录制过程，实时动态更新视频匹配结果，自动结束视频查询过程。在提高查询准确率的基础上，显著提高了用户的使用体验。为验证本文所提算法的性能，我们构建并发布了一个包含 600 小时源视频、1400 个真实查询视频的移动视频检索评测数据集。在该数据集上的评测结果显示，本文算法获得 91.59%的查询准确率，比现有最好方法提高了 4%。

3. 基于视觉-语义深度嵌入的查询相关视频缩略图选择方法

视频缩略图能集中展现视频的主要内容。而由于移动网络带宽的限制，在移动视频检索结果展示阶段选择有效的视频缩略图对于帮助移动用户快速浏览、发现目标视频起着至关重要的作用。但是现有视频缩略图选择方法没有利用视频语义信息，因而无法体现用户的查询意图。为此，本文提出了一种基于多任务学习策略的视觉语义深度嵌入方法，将不同模态的文本信息和视觉信息映射到同一个隐含语义空间，直接度量两者之间的相似度，以挖掘文本查询和视频缩略图的语义关联，使得选择的缩略图能充分反映用户的查询意图。该方法不受训练过程中语义类别的限制，能够有效应对互联网不可预测的多类别查询和视频。且通过多任务深度学习技术，该方法可在大规模带有用户点击信息的视频/图像数据集上充分挖掘用户查询与视频缩略图之间的关系。在亚马逊标注平台上 191 个标注人员参与标注的 17,480 个查询-缩略图集合上的评测结果显示，本文算法的缩略图选择效果比已有最好方法提升了约 6%。

关键词：移动视频检索，音/视频联合的紧凑视频签名，渐进式传输，音视频分层哈希索引，查询相关视频缩略图选择，多任务学习，视觉-语义深度嵌入

ABSTRACT

Mobile video search is a frontier research topic in video search communities. The proliferation of mobile devices is producing a new wave of applications that enable users to sense their surroundings with smart phones. While on the go, users always prefer mobile devices as their principal video search and browsing tools. The advanced built-in cameras have made video search very natural—mobile users can now discover videos by simply capturing a few seconds of what they are watching. As a result, mobile video search—searching similar or duplicate videos by a mobile device—has a huge application prospect and commercial value. Although there have been much effort on mobile video search, it is observed that most of them rely on the computing power of the cloud by simply sending the original query to server, while neglecting the increasing computing capacity of mobile client. Besides, unlike traditional desktop video search, mobile video search also has its unique challenges: 1) large aural-visual variance of query video; 2) stringent memory and computation constraints on the mobile devices; 3) network bandwidth limitation; and 4) instant search experience.

In this paper, we have designed an innovative mobile video system that represents one of the first attempts towards instant and progressive video search, by leveraging the light-weight computing capacity of mobile devices. In particular, the system is able to index large-scale video data using a new layered audio-video indexing approach in the cloud, as well as generate light weight joint audio-video signatures with progressive transmission and perform progressive search on mobile devices. Furthermore, the query-dependent video thumbnails are selected by multi-task deep visual-semantic embedding to help users quickly find the interesting videos. The main contributions of the thesis are summarized as follows:

1. Video Signature Generation and Acceleration with Progressive Transmission

For the purpose of fast computation, memory efficient, robust and highly compacted, we extract the visual hash bits with progressive transmission and landmark-based audio fingerprinting as the video and audio signatures, respectively. As the complementary nature of the audio and video signals, the proposed joint audio-video signatures are more robust to the large variance of query videos, especially for complex mobile video capturing conditions. Furthermore, a competitive hash function is used to significantly reduce the bits to transfer

from mobile to server. In particular, we propose a weighted Hessian response based progressive transmission to support varying signature scale in one second query video, which further decreases the number of video signatures needed to be transferred and searched. Consequently, we only require transmitting less than 0.88 KB/s audio and video signatures, which can reduce 33.5% retrieval latency with very little decrease in retrieval accuracy.

2. Layered Audio-Video Indexing and Progressive Query Process

Even though matching between binary codes is efficient, indexing search speed and audio video combination for a large video dataset is still a bottleneck in real-time mobile video search. To improve the effect, we propose a novel layered audio-video indexing scheme to holistically exploit the complementary nature of audio and video signals for more robust video search. The index effectively employs the hierarchical decomposition strategy to improve the visual points search speed, and exploits the complementary nature of audio and video signals in two fusion stages. Furthermore, we propose a progressive query process to support varying lengths in the query video, where in most cases the length is much shorter than those in any existing mobile applications. The design of the search process via a bipartite graph transformation and matching algorithm makes the video search progressive—the search can stop anytime once a confident result is achieved. This significantly improves the recognition bitrate and thus improves users' search experience. On a 600 hours video dataset, the system can outperform the state-of-the-arts by achieving 91.59% precision when the query video is less than 10 seconds.

3. Multi-Task Deep Visual-Semantic Embedding for Video Thumbnail Selection

Given the tremendous growth of online videos, video thumbnail, as the common visualization form of video content, is becoming increasingly important to influence mobile user's browsing and searching experience. In this paper, we have developed a multi-task deep visual-semantic embedding model, which can automatically select query-dependent video thumbnails according to both visual and side information. Different from most existing methods, the proposed approach employs the deep visual-semantic embedding model to directly compute the similarity between the query and video thumbnails by mapping them into a common latent semantic space, where even unseen query-thumbnail pairs can be correctly matched. In particular, we train the embedding model by exploring the large-scale and freely accessible click-through video and image data, as well as employing a multi-task learning strategy to holistically exploit the query-thumbnail relevance from these two highly

related datasets. Finally, a thumbnail is selected by fusing both the representative and query relevance scores. The evaluations on 1,000 query-thumbnail dataset labeled by 191 workers in Amazon Mechanical Turk have demonstrated the effectiveness of our proposed method.

Key Words: Mobile Video Search, Compacted Audio-Video Signature, Progressive Transmission, Layered Audio-Video Indexing, Query-Dependent Video Thumbnail Selection, Multi-Task Learning, Deep Visual-Semantic Embedding

目 录

图目录

表目录

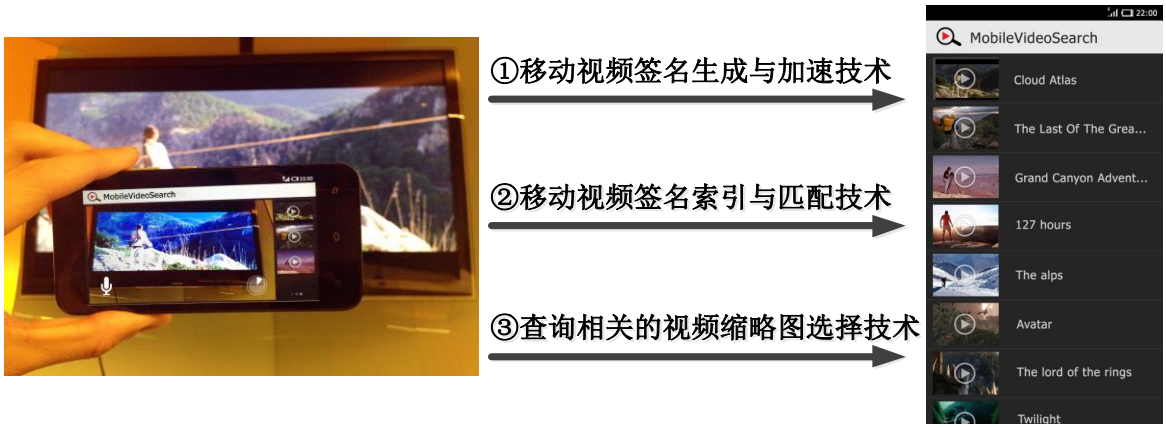
第1章 绪论

1.1 研究背景及意义

移动视频检索（Mobile Video Search）是指借助移动设备，通过输入文字查询或录制一段视频内容的方式，获取目标视频的近似视频的检索技术。在本定义中，移动设备是一种小型的、便于携带的计算设备。典型的移动设备带有一个可触摸式的显示屏、摄像头、麦克风和一块可持续电源，带有操作系统并能随时随地访问互联网，并且重量通常小于 0.91KG。常见的移动设备有智能手机、平板电脑和个人数码助手等。近似视频检索（Near-duplicate Video Search）是指在一个数据集合中找到一对/一组内容相似，但是由于录制环境、格式转换或者编辑处理而导致视频外观不同的近似视频。移动视频检索具有移动化、网络化和智能化等特点，能够为用户提供“所见即所知”的新一代视频检索服务，在推动国家信息化建设的同时，能够为移动社交网络、移动互联网电子商务、实景导航和增强现实交互等诸多实际应用提供关键技术支持。

随着多媒体技术的发展，互联网视频规模呈现了爆炸式增长。2014 年，在视频网站“YouTube”上，每天新上传的视频规模超过 10 万个小时；在社交网站“Facebook”上，每天也会有超过 10 万个新视频诞生[1]。面对如此多的视频，必须有一种快速有效的视频检索方法来帮助视频用户快速发现目标视频。据统计，近年来，我国移动互联网用户规模呈现迅猛增长态势，用户数已从 2010 年 4 月的 4.33 亿户上升至 2013 年 4 月的 8.08 亿户[2]；到 2015 年全球使用移动互联网的人数将超过桌面互联网[3]。移动设备和移动互联网用户的爆炸式增长改变了互联网上视频内容的产生、以及人们的检索和观看视频的方式，越来越多的人希望在移动中接入互联网进行视频搜索，便捷地获取全面的多媒体信息及服务。与桌面计算机相比，移动设备的便携性和无处不在的网络接入能力使其逐渐成为主要的视频访问和查询入口。而移动设备上多种先进的传感器所带来的多模态信息输入，也为视频检索带来了极大的便利。如图 1-1 所示，用户只要简单拍摄一段感兴趣的视频，即可立即进行视频检索。因此，**移动视频检索具有巨大的研究价值和市场需求。**可以预见，作为支撑未来移动互联网应用最重要的基础技术，移动视频搜索必将成为移动互联网的新一代技术突破点和利润增长点，在提升用户视频搜索体验的同时，带来巨大的经济效益和市场价值，具有广阔的应用前景。

尽管很多的研究者已经开始进行与移动视觉检索相关的研究工作（例如面向移动设备的地标识别[4]，商品检索[5]和增强现实[6]等），但是针对移动视频检索进行的研究仍然较少，移动视频检索仍然是一个极具挑战性的问题。虽然面向桌面计算机的传统视频



(a) 移动视频检索：查询输入 **(b) 移动视频检索：结果显示**

图 1-1 移动视频检索示意图

搜索受到了学术界和工业界的广泛关注和研发努力，已经积累了大量算法和技术方案可供选择[11-17]，但是移动视频搜索对传统视频检索技术提出了更高更新的挑战和要求，具有重要的科学研究价值。移动视频检索与传统近似视频检索的区别主要体现在：1）**查询视频存在严重图像形变和音频噪声**——由于不可控制的视频拍摄环境，与传统近似视频检索不同，移动视频检索中用户录制的查询视频会受到严重的噪声干扰和图像形变。因此，必须设计一种鲁棒的、区分能力强的视频签名来应对这些干扰。视频签名(Video Signature)是对一个视频对象的感知特征或简短的摘要，它将任意大小的视频对象映射到很短的位串，使得相似的视频具有相似的签名，不同的视频生成不同的签名[7]。2）**移动设备计算性能限制**——由于移动设备的CPU和内存性能仍然无法与桌面计算机相比，目前广泛使用的传统视频检索特征，由于内存占用率高或者计算复杂度高，并不能直接应用在移动设备上。3）**移动网络带宽限制——**由于检索过程是在云端进行，而目前移动设备大多通过低带宽的移动网络接入互联网，并且流量费用较高，因此我们必须充分借助移动设备的计算性能，减少网络传输的数据量。4）**查询视频长度短且为渐进式产生**——由于用户录制时间的原因，移动视频检索中的查询视频往往长度较短，只有几秒钟；不仅如此，移动视频检索中的查询视频是伴随着录制过程以秒为单位渐进式到达服务端。这就造成了移动视频检索是一个渐进式的子序列匹配问题。5）**移动用户搜索体验**——由于移动设备的特点，决定其使用者比桌面计算机用户更关注使用体验，我们必须提供一种快速、有效且具人性化的搜索过程，并且通过有效的结果展示方法帮助用户快速发现目标视频。这就要求移动视频检索与传统视频检索相比搜索过程更加快速，搜索结果更加准确。

由于移动视频检索巨大的市场需求，在工业界，目前已有一些移动应用尝试提供视

频检索功能。例如，移动视频检索应用“IntoNow”[1]和“VideoSurf”[2]分别被雅虎和微软以 2700 万美元和 7000 万美元的价格收购。其中，“IntoNow”通过使用手机录制一段大约 12 秒的音频片段作为查询输入。该应用使用“音频指纹”（Audio Fingerprinting）作为视频签名，在大规模的音频数据中快速发现目标视频。尽管如此，由于只依靠音频信息，该应用在声音嘈杂的环境中或者查询视频没有声音的情况下无法工作。另一款应用“VideoSurf”提供了基于视觉信息的移动视频检索方法。该应用需要录制一段 10 秒钟的视频片段作为查询并将其传输到服务端进行检索。但是由于该方法需要传输原始图像信息，导致其检索延迟和流量占用都非常大。同时，仅使用视觉信息也使其受拍摄环境的干扰较大。与现有方法不同，本文所提出的移动视频检索技术通过音/视频融合的方法来提高视频检索的准确率，并借助移动设备的计算性能来减少需要传输的特征数量，减少网络传输延迟。

在研究领域，已有的研究工作主要集中在怎样提高移动视觉检索识别比特率（Recognition Bitrate）上[8]。这里的识别比特率是指系统相对于视觉检索的性能与需要传输的特征规模的比值。高识别比特率意味着高响应速度，低网络传输流量，低电量消耗等优势，这些都是影响移动视觉检索的重要指标。目前在移动图像检索领域提高识别比特率的方法主要分为四大类：传输压缩过后的图像（Google Goggles[3], Kooaba[4]等），传输原始视觉特征[9] [10] [11]，传输压缩后的特征[1] [12]和传输特征的哈希签名[5] [13]。尽管这些方法能够在移动图像检索上取得较高的识别比特率，但并没有针对视频检索进行优化。而与图像相比，视频包含更丰富的模态信息，如声音、时序信息和字幕等。由于不同模态之间具有较强的互补性，研究表明，利用多模态数据进行视频检索，可以取得比过去单一模态更好的检索性能[14][15]。同时，移动设备带有的多种传感器（摄像头、麦克风、位置传感器等），也使其在多模态信息获取上具有先天优势。尽管如此，借助多模态信息进行移动视频检索也同时增加了计算开销，以及需要传输和索引的特征规模，造成了更多的电量消耗和传输延迟。因此，怎样通过多模态之间的互补性来提高视频检索的识别比特率，值得进一步研究。

在近似视频检索领域，根据使用的检索数据类型的不同，可以分为三类方法：基于音频的近似视频检索[16][17][18]，基于视觉的近似视频检索[19][20][21], 以及音/视频融合的近似视频检索[22][23]。但是，这些近似视频检索工作主要专注于提高视频检索的准确率，并没有考虑如何提高视频检索的识别比特率，也忽视了移动视频检索所面临的挑战, 如移动设备计算性能限制、移动网络带宽限制和特有的搜索体验等。

[1] “IntoNow,” http://www.intonow.com/.

[2] “VideoSurf,” http://www.videosurf.com/mobile.

[3] “Google Goggles,” http://www.google.com/mobile/goggles/.

[4] “Kooaba,” http://www.kooaba.com.

综上所述，移动视频检索是一个新兴的研究课题，具有巨大的研究价值和市场需求，而现有的研究方法并没有解决其面临的问题。针对上述问题，本文对移动视频检索相关关键技术进行了深入研究，旨在解决如下**科学问题**：针对移动视频检索的特有挑战，通过文本、音频、视觉和语义等多模态信息的融合，研究移动视频检索中的视频签名生成与加速技术、高效的索引和匹配技术，以及视频缩略图选择技术，以帮助提高移动视频检索的速度和准确率，提升移动用户的检索体验。研究移动视频检索关键技术具有以下三方面的意义：

1. 推动视频签名生成与加速技术的发展，提高移动视频检索性能。

通过研究多模态的紧凑视频签名生成和基于 Hessian 显著度加权融合的视频签名渐近式传输，在提高视频签名的辨识能力和鲁棒性的基础上，降低需要传输数据量，从而满足移动视频检索对视频签名生成与传输的特殊需求。

2. 推动二进制哈希索引与匹配技术的发展，提高移动视频检索效率。

通过研究融合音/视频的分层式哈希索引技术，探索多特征模态信息与特征索引技术的结合，提高现有二进制哈希索引的检索速度；结合渐进式视频匹配算法，解决视频子序列的高效匹配问题，进一步提高移动视频检索的速度和精度。

3. 完善多视角嵌入技术的理论研究，提高移动视频检索的用户体验。

通过研究基于多任务学习的视觉-语义深度嵌入技术，推动多视角嵌入技术的发展，探索了一种利用大规模数据进行多种任务协同训练，提高视频语义与视觉内容关联性挖掘的新途径。同时，通过基于视觉-语义深度嵌入技术的查询相关视频缩略图选择，为移动视频检索用户提供个性化的检索结果展示，提高用户的搜索体验。

除了学术上的意义，本课题的研究还可以有以下一些直接的应用：

1. 为用户提供“所见即所知”的新一代视频检索服务，充分挖掘移动设备作为视频访问入口的优势，带来更人性化的视频搜索、观看体验。
2. 在推动国家信息化建设的同时，能够为移动社交网络、移动互联网电子商务、实景导航和增强现实交互等诸多实际应用提供关键技术支持。

1.2 研究内容

本课题的研究目标是解决移动视频检索中移动设备计算性能限制、移动网络带宽限制、查询视频受外界干扰严重、快速的搜索体验等特有的挑战，通过研究适合移动视频检索的轻量级的视频签名生成与加速技术，获取高识别比特率的视频签名；进而通过融合音/视频信息的分层哈希索引与匹配技术，提高移动视频检索的速度和精度；最后通过查询相关的视频缩略图生成技术，帮助用户快速定位目标视频，建立一个完整的移动视频检索框架，为解决移动视频检索中的几个基本问题提供新途径。本文针对移动视频检

索中所涉及到的关键问题展开了深入研究，希望在推动移动视频检索关键技术发展的同时，为高速高精度的移动视频检索走向应用提供技术支持。本文的研究框架包括以下三个主要的研究内容：

1. 移动签名生成与加速技术

由于移动视频检索所面临的特殊挑战，传统视频签名技术识别比特率较低，已经无法满足移动视频检索的需求。因此，生成和传输具有高识别比特率的视频签名就成了本文研究的第一个重点。为了生成鲁棒性高、轻量级、易传输、易索引的视频签名，我们借助音/视频信息的互补性，分别提取了视觉哈希码和音频指纹特征。为了提升视觉哈希码的区分能力和压缩率，我们比较了不同的二进制编码方法。同时，针对视觉哈希码提取和传输时间复杂度仍然较高的问题，我们研究设计了基于 Hessian 显著度加权融合的渐近式传输方法。最后，经过基于几何信息的一致性验证，进一步提高了视觉签名的鲁棒性。为验证本文所提方法有效性，我们构建并发布了一个包含 600 小时源视频、1400 个真实查询视频的移动视频检索评测数据集。在该数据集上的实验验证，本文算法每秒只需要传输 0.88KB 的特征，与已有最好方法相比，减少了 33.5%的查询延迟。

2. 音/视频分层哈希索引与匹配技术

视频签名索引与匹配技术是决定移动视频检索性能的关键。然而传统的签名索引技术只针对单模态的视频签名进行索引，没有发挥多种模态信息融合的优势。针对该问题，本文研究设计了音/视频分层哈希索引，旨在研究充分挖掘不同模态特征间互补性的同时，通过分层过滤策略来提高视频签名检索的速度。除此之外，针对移动视频检索中视频匹配过程是子序列匹配的特点，我们还研究设计了基于二分图的渐进式视频匹配算法，通过二分图的转换和最大匹配，快速有效解决查询视频在源视频中的定位和相似度计算等问题，该渐进式的匹配过程，可对查询结果进行实时更新，并自动结束视频搜索过程，显著提高了用户体验。最后，我们将本文所研究的视频签名生成与加速技术、视频签名索引与匹配技术结合在一起，设计了一套完整的移动视频检索系统，并提出了具体的应用方案。经相关实验分析和用户主观易用性评价，证明了该系统的有效性和实用性。

3. 基于视觉-语义深度嵌入的查询相关视频缩略图选择方法

针对移动视频检索的结果展示问题，我们研究设计了查询相关的视频缩略图选择算法，根据用户输入的文本查询，自动生成满足用户查询意图的视频缩略图，帮助用户在移动设备上快速发现目标视频。我们首先通过研究视觉-语义深度嵌入技术，将文本查询与视频关键帧同时映射到隐含的语义空间，并在该空间中快速计算二者之间的相似度。该方法不受训练集语义类别的限制，能够有效计算真实互联网环境下，多种类别的查询与视频内容的关联性。通过结合多任务学习的策略，该方法在大规模带有用户点击信息的视频/图像数据集中充分挖掘用户查询与视频缩略图之间的联系，进行嵌入模型的训练。

最后，通过融合视觉代表性和查询相关性评分，我们选择了最能代表视频视觉和语义信息的视频缩略图。实验证明，本文算法的缩略图选择效果比已有最好方法提升了 6%，能够帮助移动用户快速了解视频内容，显著提高用户的查询体验。

1.3 论文组织

本文分别对移动视频检索中的视频签名生成与加速、视频签名索引与匹配、查询相关的视频缩略图选择等关键技术进行了深入的研究，由此构成了一个以满足移动用户视频查询准确性、多样性以及个性化查询要求为目标的移动视频检索技术研究框架。全文的内容组织如下：

第一章，论述论文的研究背景和研究意义，并概述论文的研究内容和研究成果。

第二章，对移动视频检索技术的研究现状、研究方法及存在问题进行系统的概述和总结。

第三章，阐述本文提出的融合音/视频信息的移动签名生成与加速技术。

第四章，阐述本文提出的音/视频分层哈希索引与渐进式视频匹配技术。

第五章，阐述本文提出的基于多任务学习的深度-视觉语义嵌入技术和查询相关的视频缩略图选择算法。

第六章，对论文的研究成果进行总结和讨论，并给出下一步研究工作的方向。

第2章 移动视频检索关键技术综述

2.1 引言

近似视频检索作为人们从大规模网络数据中获取目标视频的有效方法，一直是多媒体领域中的研究热点。移动视频检索技术是近似视频检索领域中一个更加前沿的研究课题，针对在移动互联网环境下借助移动设备的计算性能和多模态传感器进行有效的近似视频检索。典型的移动视觉检索框架如图 2-1 所示，主要分为视频签名提取与传输、视频签名索引与匹配、检索结果显示三个不同的阶段。

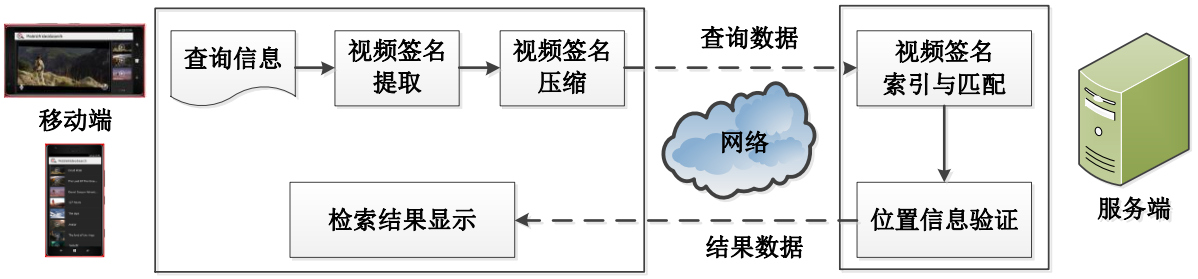


图 2-1 典型的移动视觉检索框架

移动视频检索与传统桌面近似视频检索具有一定的相似性。在传统桌面近似视频检索中，在视频签名生成阶段，对于一个给定的视频数据集，一般需要先从视频中提取底层特征，然后生成更为紧凑的、区分性更强的视频签名来代表整个视频。随着视频的快速产生，视频数据集的规模不断增加，面对大规模视频数据集，必须提供一种有效的索引结构来索引视频签名，从而提高近似视频检索的速度。视频签名生成和视频签名索引通常都是在离线阶段进行的，而视频签名匹配是一个在线过程。在该过程中，对于一个查询视频，同样需要抽取它的签名信息。之后，该签名信息被用来在源视频签名索引中检索它的近似签名。通过搜索到的近似签名，我们需要实时计算查询视频和目标视频之间的相似度。在本章中，我们将详细介绍传统桌面近似视频检索中视频签名生成，视频签名索引和匹配的相关工作，以供移动视频检索研究参考。

移动视频检索是移动视觉检索研究的一个子类。目前，很多研究机构都已经开始对移动视觉检索进行研究，其中包括 MIT、Oxford University、Stanford University、Columbia University、Google、微软亚洲研究院、北京大学、清华大学、中科院计算所和自动化所、中国科技大学、台湾国立大学和北京邮电大学等多家国内外著名大学和研究机构，并取得了一定的研究进展，研究成果发表于计算机视觉与模式识别、多媒体领域的顶级期刊（TPAMI、 IJCV、 TIP、TMM）和顶级会议（CVPR、ICCV、ECCV、DCC、ICML、ACM Multimedia）等[1] [5] [8] [9] [10] [11] [12] [13]。国际著名期刊 IEEE T-CSVT、IEEE

Signal Processing 等专门开设了针对移动视觉检索的特刊。除此之外，IEEE Mobile Multimedia Computing Workshop 也已经成功举办两届。2011 年，中科院计算所联合北京高校成立北京市移动计算与新型终端实验室，重点对移动视觉搜索进行研究并在国际顶级会议 ACM Multimedia 和国际顶级期刊 TMM 上发表了融合视觉和音频特征的移动视频搜索的文章[24][25]及相应演示系统[26]。目前，移动视觉检索中已有研究工作主要针对移动设备上的图像内容检索开展研究。由于视频可以看作是由连续图像组成的，因此现有移动视频搜索工作大多是将视频作为图像进行处理和检索。

与移动视频检索相比，移动视觉检索同样面临查询图像受拍摄环境干扰严重、移动设备计算性能限制、移动网络带宽限制和特有的搜索体验等挑战。已有的研究工作主要集中在怎样借助移动设备有限的计算性能从图像中提取高识别比特率的视觉特征作为查询[8]。高识别比特率意味着高响应速度，低网路传输流量，低电量消耗等优势。移动视频检索同样面临识别比特率的问题，但是与图像不同，视频包含更多可以使用的信息，比如文本、音频、时间信息等，如何有效利用这些信息提高移动视频检索的识别比特率，是本文研究的重点。在本章中，我们将对移动视觉检索中提高识别比特率的方法进行概述，并以此作为参考，研究提高移动视频检索的识别比特率的相关方法。

除了研究提高识别比特率，视频检索结果展示在移动视频检索中也是非常重要的一节。与图像检索结果展示完全不同，用户无法在短时间内了解检索到的视频的主要内容。而且由于移动网络带宽和流量的限制，移动视频检索用户也无法像传统视频检索中一样逐个播放结果视频来确定目标视频。因此，通过生成查询相关的视频缩略图来帮助用户了解视频内容，也成了移动视频检索结果展示中最直接有效的方法。本章中，我们对已有的视频缩略图选择算法进行综述，并分析其优势与不足，从而帮助我们研究更适合移动视频检索的视频缩略图选择算法。

2.2 移动视觉检索

在移动视觉检索的最初阶段，研究人员(Google Goggles[5], Kooaba[6]等)直接将查询图像发送到服务端，然后把特征提取、检索和匹配都放在服务端完成。这样做虽然简化了移动端的操作，但却忽略了移动设备的计算能力，受移动网络带宽的限制，直接传输图像内容会造成较大的网络传输延迟和流量损耗。近年来，更多的研究工作集中在如何在移动端提取精简的图像特征上，主要工作可以分为以下几类：

2.2.1 传统移动视觉特征

在这类方法中，研究人员直接在移动设备上提取传统图像检索中常用的视觉特征，

[5] “Google Goggles,” http://www.google.com/mobile/goggles/.

[6] “Kooaba,” http://www.kooaba.com.

如 SURF[27]和 BoW[19]等。在此基础上，研究人员对传统视觉特征的提取算法针对移动设备的特性进行了优化和改进。比如，Yang 等人[9]提出了面向移动设备的 SURF 特征加速算法。该算法使用了内容感知的图像拼接技术（Content-aware Tiling）和基于梯度矩的方向算子（Gradient Moment based Orientation Operator）。其中，内容感知的图像拼接技术将图像划分成拼接块，然后在每个拼接块上单独进行特征提取，来减少内存消耗。基于梯度矩的方向算子自动选择非均匀的拼接块大小来降低内存占用率，增大识别准确率。除此之外，中科院计算所的 Xia 等人[10]和斯坦福大学的 Chandrasekhar 等人[11]提出了图像局部特征的渐进式传输技术，通过融入特征的位置信息，只需要传输大 40%的 SURF 特征，即可获得与传输全部 SURF 特征同样的准确率。

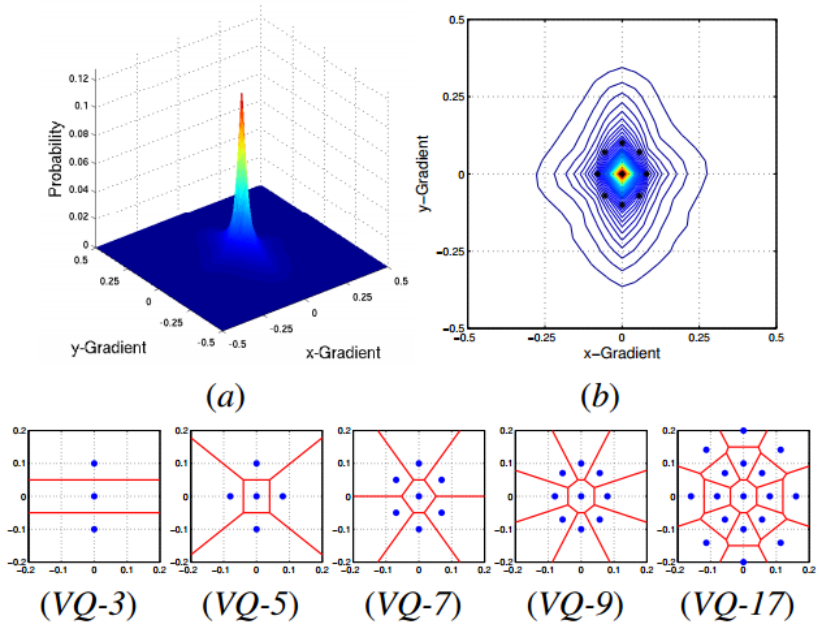


图 2-2 CHOG 特征梯度统计分布图[1]

2.2.2 移动视觉特征的压缩

该类方法通过在移动设备上对视觉特征进行压缩来减少需要传输的特征规模。其中最著名的就是 Chandrasekhar 等人[1]提出的一种压缩的梯度直方图特征（Compressed Histogram of Gradients, CHOG），该特征通过霍夫曼编码和 Gagie 树来编码梯度直方图特征，能够获得 20 倍的压缩率，大大减少了需要传输的特征数量，减少了网络传输延迟。如图 2-2 所示，该框架首先统计图像库上的特征梯度的统计学分布，然后将梯度直方图表示成树结构，并对树结构进行压缩编码，从而降低识别比特率。除此之外，他们提出了一种计算编码后的描述子之间距离的有效方法，因为不需要对描述子进行解压，从而提高查询的速度。但是，该特征在查询精度方面不如 SIFT 等图像检索领域常用的局部特征。Ji 等人[12]提出了通过多通道融合的方法来对图像检索中常用的词袋模型

（Multiple-channel Coding based compact Visual Descriptor, MCVD）进行压缩，通过融合GPS、条形码、RFID 标签等多模态信息进行提高图像检索的准确率。但是，GPS、条形码和 RFID 标签等多模态数据往往只在少数应用中存在，不能广泛使用。

2.2.3 移动视觉特征的哈希签名

除了对特征编解码压缩，He 等人[5]和 Tseng 等人[13]提出使用二进制哈希的方法对视觉特征进行编码，只需要传输特征对应的二进制码。一般该类方法分为离线和在线两个阶段。在离线阶段，算法需要在大规模图像数据集上学习哈希编码函数，尽量使编码后的二进制码能保留原有特征的区分能力。而在线阶段，只需要将视觉特征代入离线阶段学习到的哈希函数，通过与编码矩阵相乘，即可完成二进制码的转换。该方法在进行特征压缩时计算量非常小，并且二进制码在保留了原有特征区分性的基础上，压缩率高，在服务端无需再次解码，容易索引，匹配简单。这些优势使得该特征压缩方法具有很大的应用潜力。

通过以上分析可以发现，获得高识别比特率的视觉特征是移动视觉检索的目前最主要的研究目标。已有的视觉特征压缩方法，可以为移动视频检索提供非常有价值的参考。但是，目前移动视觉检索还主要针对图像进行检索。由于视频与图像相比，包含更多可用信息，如声音、时间、文本、语义等。因此，怎样将这些信息融合，获得更高的识别比特率，是本文需要研究的主要问题之一。

2.3 视频签名生成

根据视频签名所使用的数据类型的不同，已有的视频签名可以分为基于视觉的视频签名和基于音频的视频签名。

2.3.1 基于视觉的视频签名

在已有的基于视觉信息的视频签名中，研究人员通常是从各种底层视觉特征中向上生成视频签名。颜色直方图和局部特征是近似视频检索中经常使用的两种底层视觉特征。颜色直方图代表了图像中的颜色分布，它统计了图像中每种颜色值出现的频数[28][29][30]。虽然颜色直方图非常紧凑并且计算简单，但是并不包含位置、形状、纹理信息，而且对颜色改变非常敏感。不同的视频图像可能有相同的颜色分布，从而影响了颜色直方图的区分能力，不适用于移动视频检索。鉴于以上原因，局部特征成为另一种被广泛使用的底层特征。图像的局部特征描述了图像局部区域的内容信息与结构信息。目前常见的局部特征包含多种检测子和描述子[31][32][33]。图像局部特征具有较好的鲁棒性和区分性，但计算复杂度高，无法满足大规模视频数据下快速检测的需求。

鉴于移动视频检索所面对的视频数据集规模非常大，受移动设备性能限制和图像内

容严重形变的影响，怎样从底层特征中生成一种紧凑的、区分性高的视频签名来代表视频内容，成为了移动视频检索中一个主要的研究问题。根据视频签名使用的信息不同，我们将已有的视频签名生成算法分为四类：基于视频的全局签名，基于视频帧的全局签名，基于视频帧的局部签名和基于时空信息的视频签名。这四类方法主要针对了不同的近似视频类型和不同的近似视频查询任务需求，其相应的方法如表格 2-1 所示。

基于视频的全局签名使用单一的签名信息代表整个视频，该签名一般是视频全局特征分布的不同形式的表达，比如主成分、直方图或者聚类表达。常用的基于视频的全局签名有 Bounded Coordinate System [28], Accumulative HSV Histogram [29][34]和 Reference Video-Based Histogram [30]。基于视频的全局签名的主要优点是签名信息非常精简，便于存储、管理和检索。因此，基于视频的全局签名一般使用在对查询速度要求比较高的情况。但是由于该签名没有考虑到视频的局部信息，导致其区分性不高，特别是对颜色分布相似但内容不同的视频区分性非常差，不适用于移动视频检索。

表格 2-1 常见的基于视觉信息的视频签名

视频签名类型	已有工作
基于整个视频的全局签名：使用单一的视频签名代表整个视频。	Bounded Coordinate System [28] Accumulative HSV Histogram [29] Reference Video-Based Histogram [30]
基于视频帧的全局签名：使用单一的视频签名代表一个视频帧的全局信息。	Bag-of-Words [35] Glocal Descriptor [36]
基于视频帧的局部签名：使用多条视频签名代表一个视频帧的局部信息。	Local Keypoint Descriptor [31][32][33]
基于时空信息的视频签名：用时间和空间的信息与底层特征结合来代表整个视频。	CE and LBP-based Spatio-temporal Signature [20] Spatio-temporal Post Filtering [37] Video Distance Trajectory [38] Video Sketch [39]

基于视频帧的局部签名是目前最常用的视频签名。该签名使用局部特征来描述每帧图像上的局部视觉信息[31][32][33]。因为局部特征通常具有尺度、旋转和仿射不变性，由局部特征生成的视频签名能够有效应对近似视频的颜色、拍摄视角等变化。但是，相比使用全局特征，局部特征计算复杂度比较高。因此，目前基于视频帧的局部签名主要适用于对查询速度要求不高，近似视频视角变化比较多的情况。

基于视频帧的全局签名综合考虑了近似视频查询的速度和准确率。该签名生成算法使用一条签名信息来描述整个视频帧。这里的全局签名，既可以由全局特征生成，也可以由局部特征生成。常用的基于视频帧的全局签名有 Bag-of-Words [19]，Glocal Descriptor[36]等。相对于基于视频帧的局部签名，它的特征更加紧凑，提取和查询速度更快；相对于基于视频层的全局签名，该签名的区分性更强。

基于时空信息的视频签名使用空间和时间信息来描述视频。与全局签名相比，基于时空信息的视频签名由于加入了时间信息，其区分性更强。因此，近些年来，该签名在近似视频检索领域中被广泛使用[20][37] [38][39]。基于时空信息的视频签名生成算法更多的关注帧的改变、像素的运动或者兴趣点的运动轨迹，通过跟踪视频内容随时间轴的改变来查找近似视频。与其他签名相比，该签名对视频颜色和内容的改变更加鲁棒。常见的基于时空信息的视频签名生成算法有 CE and LBP-based Spatio-temporal Signature [20]，Spatio-temporal Post Filtering [37]，Video Distance Trajectory [38]和 Video Sketch [39]。但是，该算法对查询视频时间的改变特别敏感，同时，在签名匹配阶段的复杂性较高。

2.3.2 基于音频的视频签名

音频指纹是关于一段音频重要特征的紧致数字表达，它能在一定程度上反映音频的听觉质量，因此能用来有效地进行基于内容的音频检索[40]。已有音频指纹可以分为以下几类：

1. 基于 BFCC 的音频指纹

基于 BFCC 的音频指纹算法最早由 Haitsma 等人在文献[41]中提出。该算法将视频帧进行频率域转换后，在频域上按照人耳听觉的特性，将频域带进行 Bark 域频带划分，提取 32 条子带信息，得到 32 维的特征向量。然后将其按照哈希函数转换为长度为 32 的二进制哈希码。在基于 BFCC 的音频指纹基础上，Liu 等人[17]将一段音频分割为细小的片段并分别提取 MFCCs 和 RASTA-PLP 特征来代表该片段。之后，借助文本检索中词袋模型的思想，如图 2-3 所示，Liu 等人将这些分散的特征片段训练成不同的音频关键词，并将这些音频关键词的分布直方图作为音频签名使用。基于 BFCC 的音频指纹的优点在于鲁棒性高，能有效应对低码率压缩编码、全通滤波、带通滤波、重采样、均衡化、加性噪声等信号处理的影响。但是，该算法生成的指纹密度大，需要传输的特征规模大。除此之外，其索引空间大，并且检索过程涉及两段音频的指纹逐帧滑动匹配，计算量大。因此，该音频签名并不适用于移动视频检索。

2. 基于小波变换的音频指纹

基于小波变换的音频指纹算法最早由 Baluja 等人在文献[42]中提出。该算法将一维音频信号的频谱图当作二维图像进行多分辨率的 Haar 小波分解，并保留前 t 个小波系数，进行二进制编码，形成视频签名。具体的算法流程如图 2-4 所示。基于小波变换的音频

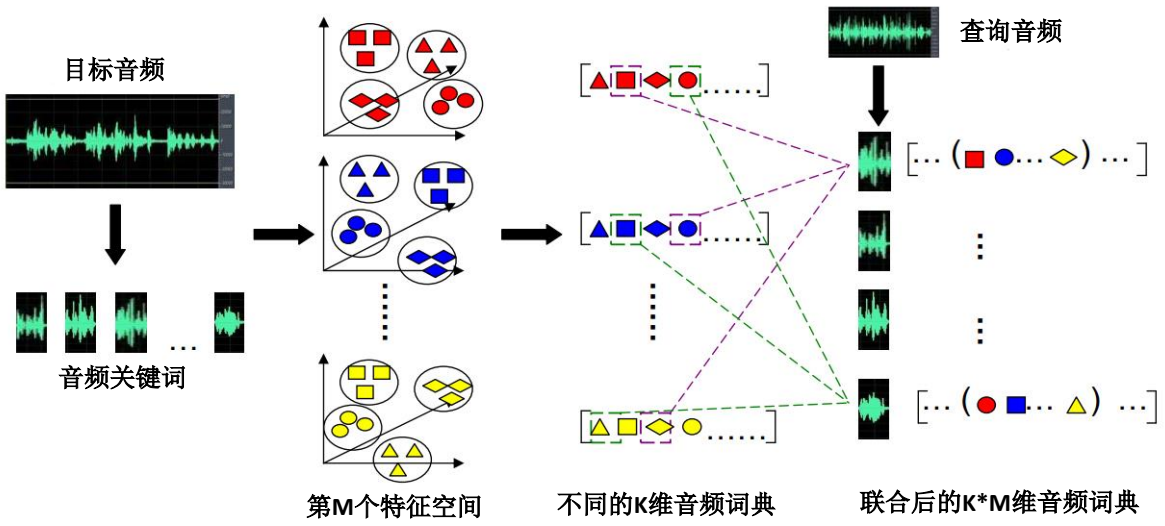


图 2-3 音频关键词模型示意图[17]

指纹区分性强，但是对于一段 10 秒钟的音频片段，仍然要产生大约 87 个长度为 100 字节的指纹特征，数据量依然较大，不适用于移动视频检索。

图 2-4 基于小波变换的音频指纹提取算法流程图

3. 基于频谱峰的音频指纹算法

在音频频域特征的基础上，Wang 等人[16]提出了一种基于频谱峰的音频签名。该特征首先计算音频帧的频率域，然后在频率域上选择局部极值点作为频谱峰，之后将多个频谱峰联合作为音频签名。该方法保证了所提取的音频签名具有较高的鲁棒性和区分性。基于频谱峰的音频指纹算法的优点在于指纹密度小、检索速度快、可扩展性强。因此，该算法在 Shazam 公司推出的音乐识别软件 Shazam 上取得了很好的应用效果[43]。

虽然音频签名具有较强的区分性，但是查询视频不一定具有有效的音频信息。而且很多视频往往具有相似的音频信息，比如无背景音等。相对于音频相似性，近似视频检索更关注视频的视觉相似性。因此，怎样将音频信息与视觉信息融合，辅助视觉信息检索，是移动视频检索研究的重点。

2.3.3 小结

已有研究[14][15]表明，利用多模态信息融合进行视频检索，可以取得比过去单一模态更好的效果。因此，多模态（Multi-Mode）[44]信息融合的近似视频检索成为视频检索领域近年来研究的重点。

现有的音/视频信息融合的方法基本上分为两种：前融合和后融合。前融合是指直接将两种签名连接变成一种更长的签名信息，该方法虽然简单，但是缺乏基本的物理解释。

由于音频信息和视觉信息分布于两个不同的向量空间，直接连接往往破坏了两种签名的内部结构。后融合是指分别使用视觉信息和音频信息进行近似视频检索，然后通过查询结果的加权融合得到优于任何单一检索的查询结果。但是，由于音/视频的度量关系并不一致，简单的加权融合无法充分发挥二者的作用。

虽然音/视频的融合可以显著提高移动视频检索的效果，但是已有的音/视频融合方法并不能充分挖掘音/视频信息的互补性。除此之外，使用两种信息进行移动视频检索，无疑会对视频签名提取、网络传输和检索带来很多的计算量。因此，移动视频检索中视频签名生成的研究重点，在于如何有效融合音/视频信息，充分挖掘音/视频互补性，生成区分性更强的视频签名。同时对音/视频融合的视频签名进行有效压缩，减少需要传输的特征规模，取得更高的识别比特率。

2.4 视频签名索引与匹配

视频签名索引的目的是构造特有的结构来存储视频签名，从而提高近似视频检索的速度。索引结构一般根据签名的类型和距离度量方法而改变。在近似视频检索领域，高维索引结构被用来应对各种视频签名的快速检索，很多索引技术非常依赖于签名的形式。一般地，多媒体索引与最近邻搜索的方法非常相关。目前常用的搜索方法有基于树结构的多维索引方法，近似最近邻方法和二进制哈希算法。

2.4.1 基于树结构的多维索引方法

最初的面向多维索引的索引都是树型结构，通过对数据空间的递归划分，将空间中相邻的数据尽可能存储到同一结点中，从而建立层次化的组织结构。按照划分方法的不同，树型结构可以分为：1. 按数据划分的索引结果，比如 R 树[45]、R+树[46]、R*树[47]；2. 按空间划分的索引结构，比如 KD 树（K-Dimensional Tree）[48]等；3. 混合型划分的索引结果（同时使用以上两种划分方式）。虽然树形索引结构实现了最近邻的精确查询和方位查询，但是特征维数超过 10 维时，该索引结构会由于空间重叠而导致索引性能急剧退化，即“维数灾难”问题。而移动视频检索中视频签名维度往往远远大于 10 维，因此树形索引结果并不适用于移动视频检索。

2.4.2 近似最近邻方法

近似最近邻 ANN（Approximate Nearest Neighbor）的方法是提高高维数据空间查询效率的常用方法。该方法通过牺牲查询的精度，换取了查询效率的提高。比如文献[49]提出的 KD 树结构，通过在查询过程中引入近似因子，增大剪枝范围，加快了树结构的搜索速度。但是，该方法仍然没有解决“维数灾难”的问题。

局部敏感哈希算法（Locality-Sensitive Hashing，简称 LSH）[50]的提出，为解决维

数灾难开辟了一个新的方向，产生了一系列基于 LSH 算法的高维索引技术[51][52][53]。LSH 类算法根据特征表述计算哈希值，并保证相似特征有较高的概率得到相同的哈希值。通过哈希值预先进行过滤，然后再在一个较小的候选集合上对特征进行线性搜索，确定最近邻元素。哈希值可以过滤掉绝大部分不相似元素，因此 LSH 类算法具有很高的检索速度。与其他的近似最近邻搜索算法相比，LSH 的查询代价达到了亚线性，并在概率上保证了找到精确最近邻的可能性。但该类方法最大的缺点就在于内存占用大，由于该算法的本质就是用空间换时间。为保证精度，相关方法的内存消耗往往 10 倍于原始特征，阻碍了它在大规模视觉搜索领域中的应用和推广。

2.4.3 二进制哈希算法

在以上介绍的近似最近邻查询方法中，经过索引过滤后的候选数据集通常根据原始距离进行排序，而在海量数据库下访问原始数据显然无法满足空间资源的要求[54]。于是，大规模视频检索下的高维索引出现了新的研究趋势，即将高维向量嵌入到压缩编码空间，并根据其压缩编码进行相似性查询。基于二进制码的索引方法把高维数据嵌入到海明空间，即映射为二进制码，使得相似的数据具有相同的编码。查询时将二进制码之间的海明距离作为距离度量方式。由于二进制码之间的比较操作可以由硬件的异或实现，因此可以实现快速查询。

谱哈希(Spectral Hashing)[55]的方法将基于谱聚类的最优化问题做了简化，并采用了一维空间下特征函数的研究成果。该方法先采用主成分分析对数据降维，然后在每一维上用特征函数对数据进行划分，进而产生二进制码。该方法的实现相对简单，并且在性能上优于其他方法。谱哈希的哈希函数学习过程是无监督的，因此最终得到的二进制编码不能反应原始数据间的语义相关性。为了解决这个问题，很多半监督 (Semi-Supervised) 和全监督 (Supervised) 的二进制哈希方法被提出[56][57][58]。比如球形哈希的方法将谱哈希中的超平面替换为超球面（Hyper-sphere）进行划分[59]。而迭代量化的方法通过学习一个正交矩阵来对谱哈希的投影矩阵进行旋转，从而降低将数据量化为二进制编码后的量化误差。图哈希使用了和谱哈希类似的策略来学习哈希函数，不同的是在学习过程中，其数据邻接关系矩阵 (Affinity Matrix) 采用了低秩矩阵乘积近似的方法构建[60]。KMH 首先对数据进行 K-means 聚类，然后使用聚类中心点的距离来代替原始数据的之间的距离[61]

二进制哈希方法在查询时使用海明距离作为距离度量，由于海明距离的取值是离散的并且受编码长度的限制，因此海明距离所能提供的距离分辨率有限，导致结果排序模糊，降低了查询结果的精度，给很多应用带来问题[62]。另外，当数据规模变大时，即使使用海明距离，查询时间依然很长。于是，[62]提出了一种查询敏感的动态排序算法 (Query-Sensitive Ranking) 对二进制哈希算法的查询结构进行精细排序。但是该算法只

能用于基于 PCAH[63]的二进制算法，其适用范围很有限。Multi-probe Hash[53]通过重复使用二进制哈希码构建索引，每次检查多个桶来提高查询精度，但该方法时间复杂度较高。Wan 等人在文献[64]中通过挖掘二进制位之间的相关性，构建自适应的多索引哈希表，可在亚线性时间复杂度内实现二进制码的精确近邻查找。但是该方法依赖于二进制码之间有很强的相关性，在实际使用中这样的二进制码并不多。为了加快查询速度，Muja 等人[65]提出了分层聚类树（Multiple Hierarchical Clustering Trees）的方法。该方法通过构建多棵索引树来提高查询速度，但同时也大大增加了内存消耗。

除此之外，目前大多数二进制哈希算法都是针对单一模态的特征进行处理，而使用单一特征进行视觉信息查询的性能是有限的。随着不同模态信息的丰富，研究从多种特征模态挖掘特征之间的关联，实现多模态检索，提高查询速度和准确率，成为值得进一步研究的重要问题。

2.4.4 视频签名匹配技术

视频签名匹配是一个在线过程。在该过程中，对于一个查询视频，首先同样抽取它的签名信息。然后，该签名信息被用来在源视频索引中搜索它的近似签名。与传统的拷贝检测[66][67][68][69]或者近似视频检索[20][70][71][72]不同，移动视频检索中的查询视频只是源视频中任意位置的一小段视频。前者虽然也是从大规模数据集中找到相似的视频片段，但是数据集中的源视频已经被切割为和查询视频相似大小的片段，只要进行相似度匹配即可。而移动视频检索是一种典型的子序列匹配任务，目标序列的长度和在源视频中的位置，都是不确定的，需要在目标中进行定位和匹配。因此，视频签名匹配首先要定位近似视频片段的位置，然后计算视频片段的相似度。目前，针对视频子序列匹配任务的视频签名匹配的方法主要分为基于滑动窗口的方法、基于路径的方法和基于图的方法[73]。

基于滑动窗口的方法定义了一个固定时间长度的滑动窗口在源视频上进行滑动，通过比较查询视频与窗口内的视频片段的相似性来计算视频匹配相似度。但是，基于固定窗口和固定阈值使得滑动窗口的方法缺乏灵活性，于是又有研究者提出了基于动态滑动窗口的方法[20]。该方法在比较查询视频与源视频比较之前，会估算待比较源视频片断的长度，动态的确定滑动窗口的大小。

由于基于滑动窗口的方法无法应对视频帧率的变化或视频帧的添加、删除等，基于路径的方法成为了视频签名匹配中常用的方法之一。文献[74]提出了一种基于树的方法用于检测视频候选子序列，然后使用剪枝策略来获取最终匹配的子序列。文献[75]则使用动态规划方法，从匹配的参考视频结果帧中查找连续的视频帧集合来计算最终的视频子序列匹配结果。

近年来，基于图的视频签名匹配方法越来越受到人们的关注。由于视频可以看作是

由连续的关键帧组成的，这些关键帧可以由图中的顶点来表示，而图中的边度量了视频帧与帧之间的相似性。这样，可以对每一对查询视频与源视频构建一个二分图，然后通过二分图最大匹配算法[76]或者网络最大流算法[77]，快速的定位视频片段并计算视频与视频之间的相似性。这种方法不受视频帧率的变化或视频帧的添加、删除的影响，并且匹配速度较快，具有很大的应用潜力。

2.5 视频缩略图选择算法

最常见的视频缩略图选择方法是选取最具视觉代表性的视频关键帧来代表整个视频。在文献[78]中，Kang 等人对视觉代表性进行了分析。他们认为用户在选择最具视觉代表性的缩略图时具有相似的倾向，并且视觉代表性跟图像质量、用户关注度检测、图像细节和播放长度有关。因此，他们使用了五种属性来衡量视觉代表性，分别是图像质量属性、图像细节属性、内容重要程度、关注度测量和镜头长度等。与以上属性不同的是，Luo 等[79]人通过摄像头的运动信息将视频分割为很多不均匀的片段，并以此来选择最具视觉代表性的视频关键帧。更进一步的，很多高层语义特征，比如目标种类、人物和主题等因素，也被用来衡量视频内容的重要程度[80][81][82]。除此之外，Lu 等人使用了基于视觉显著性的方法，通过训练一个线性递归模型来预测视频帧的视觉代表性[83]。诚然，以上这些基于视觉代表性的方法能够选择最能代表视频视觉内容的视频缩略图，但他们都忽略了视频的语义信息和用户的搜索意图等视频搜索中的重要信息，因而无法满足移动视频检索用户个性化的搜索需求[84]。

因此，近年来，更多的研究工作开始尝试选择查询相关的视频缩略图，根据不同查询的语义信息，生成更个性化的缩略图，来帮助用户更高效的找到目标视频。目前查询相关的视频缩略图选择算法的研究重点在于怎么有效计算查询语句与视频关键帧之间的相关性。根据相关性计算方法的不同，目前已有的研究工作可以分为基于图像搜索的方法和基于概念检测的方法。基于图像搜索的方法使用图片作为媒介来消除文本与视频之间的语义鸿沟 [85][86][87]。该类方法在接收到用户的查询后，直接使用这些查询去图像搜索引擎搜索相关图片，然后计算视频中关键帧与返回图像的视觉相似性，经过融合后作为查询语句与视频关键帧的相关性。尽管如此，在该类方法中，对于每个查询都要先进行图像搜索，再计算返回图像与视频关键帧的相似性，该过程会消耗大量时间，无法在实际工程中应用。于是，更多的研究人员选择使用基于概念检测的方法[88][89]。比如，Wang 等人[89]使用多示例学习的方法，将视频标签定位到视频的镜头级别，然后根据这些文本标签计算视频关键帧与查询语句之间的相关性，实现查询相关的视频缩略图选择。但是，由于多示例学习的方法无法有效处理大量类别概念的检测，该类方法只能在有限数量的查询上取得较好效果（比如文献[89]中的 60 条查询）。而在实际应用中，用户输入的查询语句是千差万别的。因此，目前基于学习的方法的瓶颈是概念检测器所

能有效检测的视觉概念的数量。由此可见，目前查询相关的缩略图选择方法的研究难点依然是如何计算查询与视频关键帧之间的相似性。

多视角嵌入（Multi-view Embedding）的方法被普遍用来解决查询语句与图像之间相似性的计算问题[90][91]。比如，Pan 等人提出了基于用户点击信息的多视角嵌入方法来计算查询文本和图像内容之间的多视角距离[92]。该方法是通过多视角嵌入来将文本和图像内容映射到隐含的子空间，并且尽可能多的保留原有空间的相似性信息。除此之外，视觉-语义深度嵌入模型从大规模文本数据中学习文本嵌入模型，将查询文本映射到隐含的语义空间，同时使用深度网络的方法，通过相似性度量学习，将图像内容也映射到同样的语义空间，从而可以在该空间中直接计算文本与图像的语义相似性[93]。这两种方法都有一个共同的优点，可以处理大规模的文本/视觉概念之间的相似性。特别是视觉-语义深度嵌入模型，通过在大规模数据集上进行语义相似性的学习，对于训练集中未出现过的文本/图像内容也能进行有效的相似性计算。

虽然多视角嵌入的方法能够有效计算查询语句与图像之间的相似性，但是因为图像检索与视频缩略图选择这两个不同的任务之间的区别，我们不能直接将图像检索中的嵌入模型直接应用到视频中来。怎样利用大规模视频数据，训练得到适合进行查询相关的视频缩略图选择的视觉-语义嵌入模型，值得更进一步研究。

2.6 小结

综合以上对国内外研究现状的调查，我们有以下结论：

1. 虽然针对近似视频检索的研究工作很多，但是由于移动视频检索与传统近似视频检索的区别，并不能直接将这些研究工作用于移动视频检索。针对移动视频检索中查询视频图像发生严重形变的情况，现有主要借助视觉信息进行视频相似性判断的近似视频检索技术都不能取得较好的检索效果。除此之外，我们还需要考虑移动设备计算性能和网络带宽的限制，查询视频典型的子序列匹配问题以及移动用户特有的搜索体验等。因此，怎样在视频签名生成、视频签名索引和视频签名匹配三个阶段解决这些问题，是我们需要研究的重点。
2. 获得高识别比特率的视觉特征是移动视觉检索的目前最主要的研究目标。已有的视觉特征压缩方法，可以为移动视频检索提供非常有价值的参考。但是，目前移动视觉检索还主要针对图像进行检索。而现有移动视频检索方法，大多直接把视频看作连续的图像，通过图像匹配实现视频检索，识别率较低，无法应对移动视频检索对检索精度的需求。由于视频与图像相比，包含更多可用信息（声音、时序、文本等），怎样将这些多模态信息融合，获得更高的识别比特率，是移动视频检索需要研究的主要问题之一。
3. 作为移动视频检索的重要一环，选择更符合用户查询意图的视频缩略图作为移动视

频检索的结果展示，在帮助用户节省查询时间和网络数据流量上具有非常关键的作用。但是已有视频缩略图选择算法大多集中在选择视觉代表性强的视频关键帧上，忽视了视频的语义信息和用户的查询意图。而现有查询相关的视频缩略图选择算法又无法应对实际应用中多种多样的查询类别。因此，怎样利用视觉-语义深度嵌入方法，有效的选择符合用户查询意图的视频缩略图，从而提高用户的移动搜索体验，也是本文需要研究的主要问题之一。

第3章 移动视频签名生成与加速技术

3.1 概述

视频签名(Video Signature)是对一个视频对象的感知特征或简短的摘要。它将任意大小的视频对象映射到很短的位串，使得相似的视频具有相似的签名，不同的视频生成不同的签名。视频的签名通常具有无冲突性、安全性、紧凑性、鲁棒性和篡改敏感性等特征[7]。作为移动视频检索的第一步，视频签名的设计非常重要，通常需要满足以下要求:

1. 与传统的近似视频检索不同，由于视频录制中不可避免的外界环境干扰，移动视频检索中的查询视频与原始视频相比，往往会发生严重的改变，主要包括：1. 视频原有的音频信息可能会受到外界噪声的干扰，甚至没有捕获到音频信息；2. 如图 3-1 所示，图像内容可能发生光照变化、图像模糊、镜面反射、图像旋转和尺度变化等严重形变。因此，移动视频检索中所使用的视频签名必须对这些音视频的变化足够**鲁棒**。
2. 视频区别于与其他媒体类型（文字、图片和音频）最主要的特点就是包含丰富的音频和视觉信息。由于视觉信号和音频信号之间的互补性，**音视频融合**的视频签名，相较于单一模态的视频签名，在应对移动视频检索中查询视频的变化更加有效。
3. 由于移动设备计算性能的限制，在移动视频检索中只能使用**轻量级**，计算简单的视频签名，能够在移动设备上快速提取。
4. 由于网络带宽的限制，我们必须对这些视频签名进行压缩，使其足够精简，**易于传输**。
5. 大规模的源视频为服务端的视频签名索引与匹配带来了极大挑战，因此，移动视频检索中的视频签名必须**易于索引和快速查找**。

基于以上原因，本文分别选择基于频谱局部显著性的音频指纹特征（Landmark-Based Audio Fingerprinting）和视觉哈希码（Visual Hash Bits）作为移动视频签名。针对视觉哈希码特征提取和传输延迟仍然较大的问题，我们设计了一种基于 Hessian 显著度加权融合的渐近式视频签名传输方法。除此之外，我们还搜集并公布了世界上第一个基于真实环境录制的移动视频检索数据集，并在该数据集上评测了我们提出的移动视频签名的性能。下面，将分别介绍相关研究内容。

图 3-1 移动视频检索中查询视频图像形变示例

3.2 视觉哈希码的提取与渐进式传输算法

3.2.1 视觉哈希码的提取

如前文所述，用户在使用移动设备进行视频查询的时候，由于操作不当和外界环境的干扰，必然会使录制的查询视频发生严重形变，比如光照变化、图像模糊、镜面反射、图像旋转和尺度变化等。所以我们必须选择一种既能在移动设备上快速提取，又对这些形变足够鲁棒的视频特征。在本文工作中，我们选择了兼具高效率和高鲁棒性的 SUFT 视觉特征。根据相关的评测报告显示[7]，与目前流行的其他视觉特征相比，如 BRIEF[94], ORB[95]等，SURF 特征对于图像旋转、尺度变化、图像模糊、光照改变等干扰更加鲁棒。另外，移动视觉检索相关工作[5][9][96]也指出，SURF 特征在高性能和高鲁棒性方面的均衡表现，使其更适合作为移动视觉检索的底层特征。此外，目前也有很多针对在移动设备上 SURF 特征提取的加速算法可以应用。尽管如此，直接传输原始 SUFT 特征是非常耗费时间和网络流量的。为此，如图 3-2 所示，我们使用了两种策略来提高 SURF 特征提取和传输的速度。

视觉哈希码的提取过程如图 3-2 所示。首先，对于一段视频，我们以秒为单位，每秒抽取一帧视频图像，然后将其归一化到小于 200×200 像素。这里对图像进行归一化主要有三个原因：1. 图像缩小后提高了在移动设备上对图像进行特征提取的速度；2. 减少了需要通过网络传输的特征数量；3. 根据我们的实验验证，尺度缩小在提高特征提取和传输速度的基础上，对于视频检索准确率影响却很小。

如前文所述，尽管我们选择了兼具高性能和高鲁棒性的 SURF 特征，但是由于 SURF 特征仍然不够精简，直接传输原始的浮点数特征向量仍然会占用较高的网络带宽，造成较长的网络传输延迟。受哈希词袋模型（Bag of Hash Bits）思想的启发[5][13]，我们使用了二进制哈希算法 (Binary Hashing)对提取的 SURF 特征进行压缩。给定一个数据集，

[7] “Feature descriptor comparison report”, http://computer-vision-talks.com/articles/2011-08-19-feature-descriptor-comparison-report/.

二进制哈希算法通过一组哈希函数 (Hash Function) 将每个元素嵌入到海明空间，映射为二进制码，使得相似的数据具有相同的编码。通过这种方式，二进制哈希算法显著降低了算法对空间的需要，有着很好的存储性能。网络传输时，二进制编码与原浮点数特征向量相比，显著减少了需要传输的特征规模，提高了传输的速度。查询时，将查询映射为二进制编码，通过计算查询编码和数据库中编码之间的海明距离找出和查询相关的结果。由于二进制编码之间的比较操作可以由硬件的异或实现，所以二进制哈希算法同样有着很好的查询性能。

目前流行的二进制哈希编码方法有谱哈希[55]、球形哈希[59]、迭代量化[97]、图哈希[60]和 K-means 哈希（简称 KMH）[61]等。其中，谱哈希、球形哈希和迭代量化的方法思想比较相似。谱哈希方法先采用主成分分析对数据降维，然后在每一维上用超平面对数据进行划分，进而产生二进制码[55]。球形哈希的方法将谱哈希中的超平面替换为超球面进行划分[59]。而迭代量化的方法通过学习一个正交矩阵来对谱哈希的投影矩阵进行旋转，从而降低将数据量化为二进制编码后的量化误差。图哈希使用了和谱哈希类似的策略来学习哈希函数，不同的是在学习过程中，其数据邻接关系矩阵采用了低秩矩阵乘积近似的方法构建[60]。KMH 首先对数据进行 K-means 聚类，然后使用聚类中心点的距离来代替原始数据之间的距离[61]。由此可见，后四种哈希方法都是以谱哈希方法为基础，通过替换其中某一部分来进行改进。然而这些改进通常是针对近似图片检索，在面对大规模视频检索时，由于视频中视觉信息的冗余性，五种方法的准确度是基本相似的。为了验证该结论，我们将谱哈希方法和目前在近似图片检索中表现效果最好的 KMH 方法进行了对比，发现两种方法在移动视频检索上取得了基本一致的命中率（Hit Ratios）。具体的实验结果请参见图 3-5。在本文工作中，我们综合考虑算法性能和复杂度，选择了谱哈希算法作为对 SURF 特征进行二进制编码的方法。

谱哈希算法中的哈希函数如公式(3-1)所示：

$$h = sign(\cos(Wx)) \tag{3-1}$$

其中 x 是 SURF 的特征向量，W 是学习到的哈希矩阵，h 是转换后的二进制编码。根据文献[5][13]给出的建议参数，我们将二进制编码的长度设置为 80 比特。除此之外，我们使用 8 比特来存储 SURF 特征的主方向，用来进行位置信息验证。因此，对于每一个 SURF 特征向量，我们都将其压缩成长度为 88 比特的二进制编码。

3.2.2 视觉哈希码的渐进式传输

目前，大多数的移动设备都包含一个多线程的 CPU，因此，我们可以并行地提取视频的视觉特征和音频特征。尽管如此，根据我们的实验表明，视觉特征的提取和传输时间是音频特征的 3 倍。因此，怎样进一步减少视觉特征的提取和传输时间，是降低整个移动视觉检索时间延迟的关键。

根据斯坦福大学 Chandrasekhar[11]和中科院计算所 Xia[10]等人的研究表明，在移动图像检索中，只需要提取图片的部分特征，即可以达到与传输所有特征相似的图像检索准确率。而视频与图像相比，包含更多相似的图像帧和音频信息，因此也就有更多冗余的信息。因此，在本文提出的移动视频检索系统中，使用了基于 Hessian 显著度加权融合的视觉哈希码渐进式传输方法，即根据视觉哈希码的重要性，渐进式地进行特征传输，并在根据具体的设备性能和网络环境，自动决定需要传输的特征规模。使用该方法，可以显著降低视频签名生成与传输的时间，满足实时的移动视频检索。

视觉哈希码的渐进式传输过程如图 3-2 所示。首先，仍然是对视频图像重新调整大小，并平均分割为 9 个相同大小的区块。每个区块中的所有 SURF 特征点都可以看作是一个查询单元。之后，使用公式(3-2)计算每个单元的权重。

$$\text{score}_i = \omega_i \times \sum_{j=1}^{n_i} p_j \,, \tag{3-2}$$

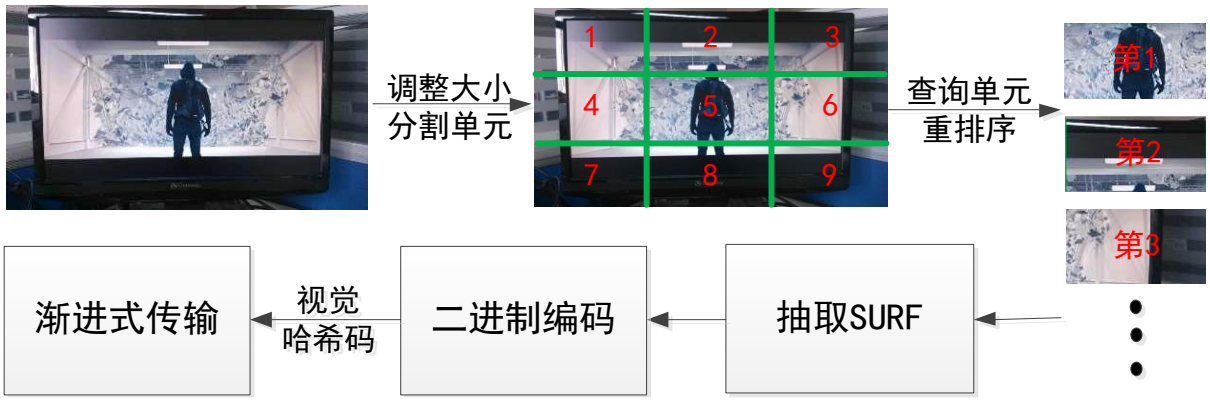


图 3-2 视觉哈希码的提取和渐进式传输过程

公式中 w_i 是每个查询单元的初始化权重。根据观察 3.5 节中真实环境下录制的查询视频数据发现，用户在录制查询视频时，习惯将视频内容放在屏幕的中心区域。因此，我们将九个单元的初始化权重分别设置为：$\omega_5 = 0.25, \omega_2 = \omega_4 = \omega_6 = \omega_8 = 0.125,$ $\omega_1 = \omega_3 = \omega_7 = \omega_9 = 0.0625.$ 公式中的 p_j 是区块中每个 SURF 特征点的 Hessian 响应值，n_i 是每个区块中的 SURF 特征点的个数。SURF 特征的提取过程中使用 Hessian 矩阵检测兴趣点，响应值越大代表该特征点拥有的信息量越大，因此需要优先传输这些特征。在 3.6 节的实验中，我们比较了 Hessian 矩阵响应值和其他显著性权重计算方法，证明了其效果最好。所有单元的权重计算完毕后，我们会按照权重的高低顺序提取和传输视觉哈希码。最后，需要传输的查询单元个数可以根据两个准则来判断：1) 选取固定个数的单元，比如根据我们的实验，传输 3 个单元既能获得与传输所有单元近似的准确率，又能获得非常显著的速度提升。2) 选取一个固定的提取和传输时间，比如 500ms，如果超过了这个时间，则立即停止视觉哈希码的提取和传输。设置固定时间的原因是可以

根据真实设备和网络环境的不同，在满足实时查询的基础上，动态地调整需要传输的特征规模。

3.3 基于频谱局部显著性的音频指纹提取算法

音频指纹是一个音频片段中重要特征的紧致数字表达，它能在一定程度上反映音频的听觉信息，因此可以有效的应用于基于内容的音频检索[40]。本文之所以选择音频指纹作为移动视频检索中视频签名的一部分，主要有以下几点原因：1）相对于音频数据本身，音频指纹的数据规模通常较小，这样可以减少需要传输的特征规模，提高传输的速度；2）音频数据生成的索引规模比较小，相应的搜索空间也就比较小，这样可以在大规模视频库上进行快速检索；3）音频指纹在很大程度上反映了音频的听觉特征，因此具有较高的鲁棒性，即使音频信号有一定程度的失真，如背景噪声、压缩编码等，仍然能被正确检索。根据音频指纹中特征提取方法的不同，目前常见的音频指纹可以分为时域音频指纹和时频率域音频指纹。因为时频域指纹能同时反映音频信号在时间和频率两个维度上的信息，目前在音频检索领域得到了广泛的应用。为了满足移动视频签名鲁棒性高、特征规模小、检索速度快和增量检索等特点，我们选择了基于频谱局部显著性的音频指纹。

如图 3-3 所示，基于频谱局部显著性的音频指纹提取过程主要分为：1. 音频数据预处理；2. 经频域转换获取频谱图；3. 频谱图上的局部显著点选取；4. 将显著点组合为音频指纹。具体过程如下：

1. 音频数据预处理

首先，我们对一秒钟的音频片段进行解码，获取单声道音频，并重采样至 8k Hz。然后，重采样后的音频片段被分割为长度为 256 毫秒，步长为 32 毫秒的音频帧。较大的帧间重叠使得相邻帧的特征具有较强的相关性，并随时间缓慢变化，使得音频指纹在一定程度上对帧移误差具有鲁棒性。

2. 经频域转换获取频谱图

傅利叶变换能将一个满足一定条件的函数分解到一个垂直的坐标系，每个坐标分量称为频率，在这个坐标系下的系数（本身是一个函数），被称为这个函数的频谱[98]。傅里叶变换是一种线性变换，可以将时域信号转换到频率域，从能够很好的刻画信号的频率特征。在本文中，我们对每段 256 毫秒的音频帧进行短时傅里叶变换，采样率设置为 8000Hz, 每帧有 2048 个采样点，频率分辨率约为 4Hz。变换之后得到的频谱图如图 3-3 所示。

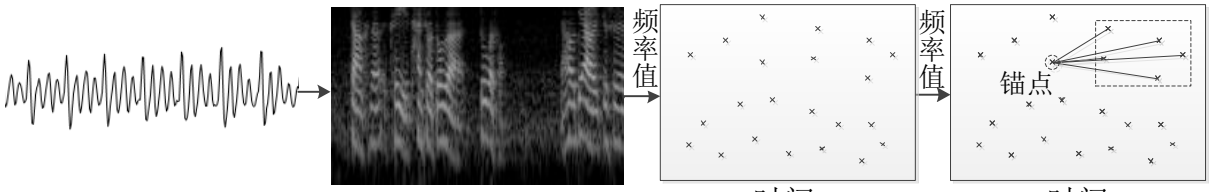
3. 频谱图上的局部显著点选取

在频域转换后得到的频谱图上，白色的点代表了音频能量聚集的点。这些聚集点对各种信号失真具有很强的鲁棒性和近似线性叠加性，适合用来做特征点。我们通过如下

准则选取显著点：1. 能量值高于周围的点；2. 振幅值高于周围的点；3. 近似符合线性分布。

4. 将显著点组合为音频指纹

由于单个显著点并不能够提供足够的信息熵来确定两段音频的相似性，于是，我们将显著点两两配对，生成信息熵更高的组合指纹。如图 3-3 所示，对于一个显著点(t_i, f_i)，将其作为锚点（Anchor Point），并在其附近选择一个矩形区域，将该显著点与矩形区域里面的每个显著点两两组合成为地标（Landmarks）。每个地标可以表示为四维的特征向量$l_i = \{t_i^a, f_i^a, \Delta t_i^a, \Delta f_i^a\}$。这里$t_i^a$和$f_i^a$是锚点的时间偏移和频率值，$\Delta t_i^a$和$\Delta f_i^a$是锚点和配对点之间的时间差和频率值的差。最后，我们使用哈希的方法将地标压缩为音频指纹$l_i = \{h_k^a, t_i^a\}$。这里h_k^a是$f_i^a, \Delta t_i^a$和Δf_i^a的哈希值，长度为 25 比特，不同的音频指纹可能含有相同的哈希值。t_i^a仍然为锚点的时间偏移，长度为 15 比特。最后，音频指纹的长度为 40 比特。对于一秒钟的音频片段，我们随机选择 100 个音频指纹，因此，一秒钟只需要传输 0.5 KB/s 的音频指纹。



(a) 音频帧　(b) 音频频谱图　(c) 候选极值点　(d) 音频签名

图 3-3 基于频谱局部显著性的音频指纹提取过程

3.4 移动视频签名的索引与匹配

3.4.1 基于分层聚类树的二进制哈希索引

当用户进行查询视频的录制时，系统会在后台实时提取视觉哈希码和音频指纹作为视频签名，并通过互联网传输到服务端进行近似视频查询。将视频签名压缩为二进制码后，在近似签名的检索时，可以直接计算签名之间的海明距离。海明距离可以直接通过位与位之间的异或操作来计算，因此匹配速度相比压缩前有了很大提高。尽管如此，如果简单地在大规模视频库上进行线性搜索，仍然非常费时费力。目前流行的解决方案是将线性搜索替换为近似匹配算法，通过构建二进制哈希索引来进行近似最近邻的查找。目前常用的索引方法有局部敏感哈希算法（Locality Sensitive Hashing）[50]、语义哈希算法（Semantic Hashing）[56]和最小哈希算法（Min-hash）[99]等。在本文中，我们选择了基于分层聚类树（Hierarchical Clustering Tree）的二进制哈希索引方法[65]。该方法使用分层过滤的思想来减少查询空间。它通过聚类将输入的数据构建为一颗树，树中每个非叶节点为一个聚类中心，叶节点为匹配到的点。分层聚类树的构建算法如表格 3-1

所示。该算法首先将数据集中的所有点聚类为K个类别，K为分支因子（Branching Factor）。聚类开始时，随机的选择 K 个点作为聚类中心，然后把其他所有点都按照与聚类中心的距离进行划分。重复整个算法直到每个类别中的点个数都小于指定的阈值，并把他们作为叶节点。

表格 3-1 分层聚类树的构建算法

输入：特征数据集D
输出：分层聚类树
参数：分支因子K，最大叶节点个数S_L
1: **if** size(D) $< S_L$ **then**
2: 将D中所有点作为叶节点
3: **else**
4: P ← 从 D 中随机选择 K 个点
5: C ← 以 P 中的点为中心将 D 中的点进行聚类
6: **for** $C_i \in C$ **do**
7: 将P_i作为非叶节点
8: 对C_i中的点递归调用该算法
9: **end for**
10: **end if**

分层聚类树的搜索过程如表格 3-2 所示。首先，对于每一颗树，算法从上往下选择与查询点近似的点并且递归的向下探索，同时将没有探索到的点加入优先队列PQ。搜索过程到达叶节点时，线性地搜索叶节点中的所有点。当每棵树都被搜索过后，如果仍未检索到足够的点，则继续从优先队列 PQ 中进行查找，直到搜索的点的个数超过L_{max}。L_{max}指定了近似搜索算法的度，L_{max}越大，搜索到的最近邻越多，但是搜索的时间代价也越大。

表格 3-2 分层聚类树的搜索算法

输入：分层递归树T_i，查询点Q
输出：查询点Q的K最近邻
参数：需要匹配的点的最大个数L_{max}
1: L ← 0 {L = 搜索的点的个数}
2: PQ ← 空优先级队列

3: R ← 空优先级队列
4: **for** 每棵树T_i **do**
5: 调用TRAVERSETREE(Ti, PQ, R)
6: **end for**
7: **while** PQ 非空 and L < L_{max} **do**
8: N ← PQ 的头
9: 调用TRAVERSETREE(N, PQ, R)
10: **end while**
11: **return** R 的前 K 个点
函数TRAVERSETREE(N, PQ, R)
1: **if** 节点 N 是叶节点 **then**
2: 搜索 N 中的所有点并添加到 R
3: L ← L + |N|
4: **else**
5: C ← N 的子节点
6: C_q ← C 中与查询点 Q 最相近的点
7: C_p ← C\C_q
8: 将C_p中的所有点加到 PQ 中
9: 调用 TRAVERSETREE(C_q, PQ, R)
end if

3.4.2 基于几何一致性验证的重排序

当得到每个视频签名的前 K 个最近邻之后，我们使用投票的方法来获得每个查询关键帧的近似关键帧。这里，在投票之前，我们会使用几何一致性验证的方法来对查询到的视觉哈希码进行重排序。移动视觉搜索中的几何一致性验证是指通过验证特征点之间的空间关系来判特征点之间是否为正确匹配的一种技术。几何一致性验证可以过滤掉很多错误匹配，提高近似关键帧检索的准确率。在本文中，我们使用了两种快速但非常有效的几何一致性验证方法[96]。第一种验证方法认为匹配到的特征点的主方向应该是一致的，即考虑到图片的旋转，所有匹配上的特征点的主方向的差应该是一致的。对于两个近似视频帧，该方法会计算每对匹配上的特征点的主方向差$\Delta\theta_d$，然后统计所有$\Delta\theta_d$的直方图分布。为了简化计算，我们将$\Delta\theta_d$归集到 10 个桶，哪个桶的统计值最高，则表明了图像旋转的角度。最后，我们只保留方向差落到该桶内的匹配点。第二种验证方法认为，正确匹配的特征点应该来自同一个区块。因此，对于查询图像每个区块中的特征点，我们需要找到在近似图像中的区块位置，即匹配上最多特征点的目标区块。这样，其他区块则会被过滤掉。这两种位置验证方法由于只用到了方向角和区块编码，无需传输特

征点的位置信息，与其他几何一致性验证方法相比，可以在保证验证性能的同时，显著节省需要传输的数据量。

3.5 移动视频检索评测数据集

虽然在移动图像检索领域早已存在被研究者公认的评测数据集（Stanford[100]，MVSBench[10]），但是在移动视频检索中，却一直没有公开评测数据集。在以有的研究工作中，研究者经常使用特定程序来对原始视频进行攻击，模拟查询视频录制过程中因外界环境干扰所发生的变化。但是，这些人为设定的变化，并不能真实客观反映在真实世界中移动设备使用者所拍摄的查询视频的情况。为了解决移动视频检索中最基本的数据问题，全面准确的研究移动视频检索与普通视频检索的区别，我们构建并发布了一个基于真实环境录制的移动视频检索数据集。

为了保证数据搜集过程的真实客观，我们邀请了 25 名志愿者，他们的年龄、职业和性别比例如表格 3-3 所示。为了保证录制质量，真实反映目标人群的使用习惯，这些志愿者都有不少于 1 年的移动设备使用经验，并且经常在移动设备观看视频。

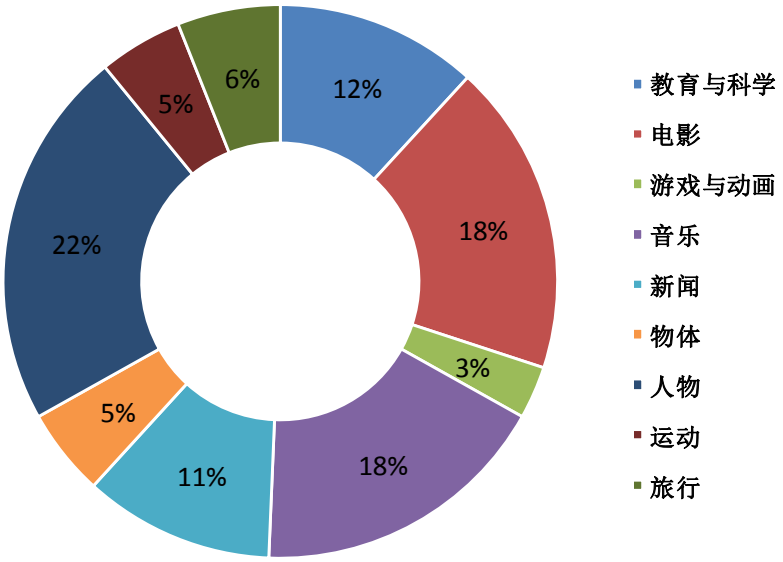


图 3-4 移动视频检索数据集中的查询视频类别分布图

之后，我们使用国际视频检索评测机构 TRECVID 2011 年的视频数据集中的 19,200 个视频作为源视频数据集[8]。这些数据来自电视和互联网，包含 9 大类别，时长共 600 个小时的视频数据。之后，志愿者被要求在这 19,200 个视频中挑选若干视频，然后使用自己的移动设备，在他们真实的生活环境中按照自己的使用习惯，录制不小于 30 秒的查询视频。他们使用的移动设备包括苹果 4S 手机、苹果 IPAD 平板电脑、HTC z710t、三星 Galaxy S3 等多种移动设备。最后，我们共获得 1,400 个长度不少于 30 秒的查询视

[8] TRECVID, http://trecvid.nist.gov/.

频。图 3-4 展示了这些查询视频的类别分布。这里的分类标准来源是 Youtube 视频网站的分类标准，共包含教育与科学、电影、游戏与动画、音乐、新闻、物体、人物、运动和旅行 9 大类。目前，该视频数据集已公开发布[9]，作为移动视频检索领域的基准评测数据集，被台湾国立大学、微软亚洲研究院、中国科学技术大学和北京邮电大学等多家研究机构使用。

表格 3-3 查询视频录制人员的年龄、职业和性别比例

年龄	比例	职业	比例	性别	比例
21~30	56%	Student	56%	Man	68%
31~40	32%	Staff	36%	Women	32%
41~50	12%	Manager	8%	—	—

3.6 实验结果

本章节将对移动视频签名生成与加速技术进行综合的评测，包括不同二进制编码方法的评测、视觉哈希码渐进式传输的评测和移动视频检索准确率的评测。所有的评测，都在 3.5 节中介绍的移动视频检索评测数据集中完成，共包含 1400 个查询视频和 19,200 个，600 小时规模的源视频。为了更好地评测不同的方法，我们使用命中率（Hit Ratio）作为评价指标。这里，命中率是指随着查询时间的增加，在前 K 个返回视频中，成功找到目标视频的查询占总查询次数的百分比。由于系统可以返回给用户多个匹配视频，因此，K 分别取 1, 5 和 10。

3.6.1 不同二进制编码方法性能评测

首先，我们评测了视觉哈希码生成中，使用不同哈希函数进行二进制编码的准确率。这里，总共比较了谱哈希(简称 SH)和 K-means 哈希（简称 KMH）两种二进制编码方法。其中，谱哈希是我们最终使用的方法，K-means 哈希是目前在图像检索应用中性能最高的方法。KMH 方法使用了作者提供的代码和推荐的参数配置。为了客观评价，我们设计了如下 4 中评测方案：1. 基于谱哈希的视觉哈希码方法（VBM-SH）；2. 基于 K-means 哈希的视觉哈希码方法（VBM-KMH）；3. 基于谱哈希的视觉哈希码与音频指纹融合的方法 （LAVE-SH）；4.基于 K-means 哈希的视觉哈希码与音频指纹融合的方法（LAVE-KMH）。最后，我们在移动视频检索数据集上测试了这四种方案在前 10 个返回

[9] “移动视频检索评测数据集”，http://mcg.ict.ac.cn/mcg-mvq.html

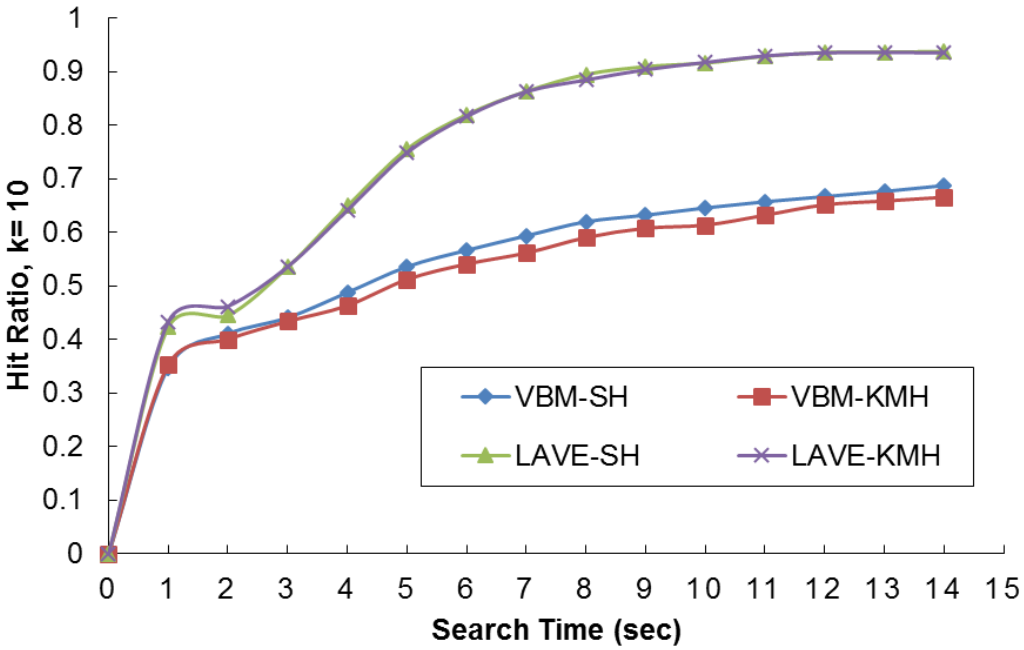


图 3-5 视频签名生成中不同二进制编码方法的性能比较

视频上的命中率。实验结果如图 3-5 所示。从图中可以看到，无论是只使用视觉哈希码，还是与音频指纹进行融合，两种二进制编码方法都取得了近似的查询准确率，其中谱哈希方法的准确率稍高一些。该结果证明，与图像不同，视频多个帧间图像包含了更多的冗余信息，较为简单的谱哈希方法既能取得较好的压缩效果。而且由于谱哈希方法的计算复杂度更低，因此我们最终选择了谱哈希作为移动视频检索中视觉哈希码的二进制编码方法。尽管如此，我们的系统并不依赖于某个特定的二进制编码方法，任何更为有效的二进制编码方法都可以简单方便的替换现有编码方法。

3.6.2 移动视频签名提取和传输时间评测

3.6.2.1 移动客户端视频签名提取时间

我们在两台不同的移动设备上测试了视频签名抽取的时间。设备 I 使用 Android 4.1.1 操作系统，搭载了一颗高通 APQ8064P，主频为 1.5GHz 的处理器，2GB 内存和 16GB 存储空间。设备 II 使用 Android 4.2.2 操作系统，搭载了一颗 Texas Instruments OMAP 4460 1.2 GHz 主频的处理器，1GB 内存和 16GB 存储空间。对于一秒钟的视频片段，在设备 I 上提取音频签名需要 120ms，提取所有视觉哈希码需要 261ms；在设备 II 上提取音频签名需要 130ms，提取所有视觉哈希码需要 272ms。作为对比，我们实现了文献[22]中的视觉特征提取算法，该算法对 SURF 特征进行了无损编码。该算法在设备 I 和设备 II 上的时间分别为 267ms 和 280ms。而经典移动视觉签名 CHOG 的提取时间则为 1000ms[1]。

3.6.2.2 视频签名网络传输时间

因为视频签名网络传输时间非常依赖测试时的网络环境，所以我们使用每秒需要传输的视频签名的比特率来评价传输时间。如第三章介绍，我们将视觉特征通过哈希函数压缩为 88 比特的哈希码，每秒钟大约提取 52 个局部特征点。而对于音频指纹，我们同样压缩为 40 比特的二进制码，每秒钟提取 100 个音频指纹。因此，每秒钟只需要传输小于 1.125KB/s 的融合音视频的视频签名。如果采用渐进式传输策略，只传输前三个特征单元，则只需要传输 0.875KB/s 的视频签名，远小于 CHOG（4KB/s）等经典移动视觉签名。理论上，对于一个典型的带宽为 40KB/s 的 3G 网络，传输所有视频签名只需要 29ms。

3.6.3 基于 Hessian 显著度加权融合的渐近式传输方法性能评测

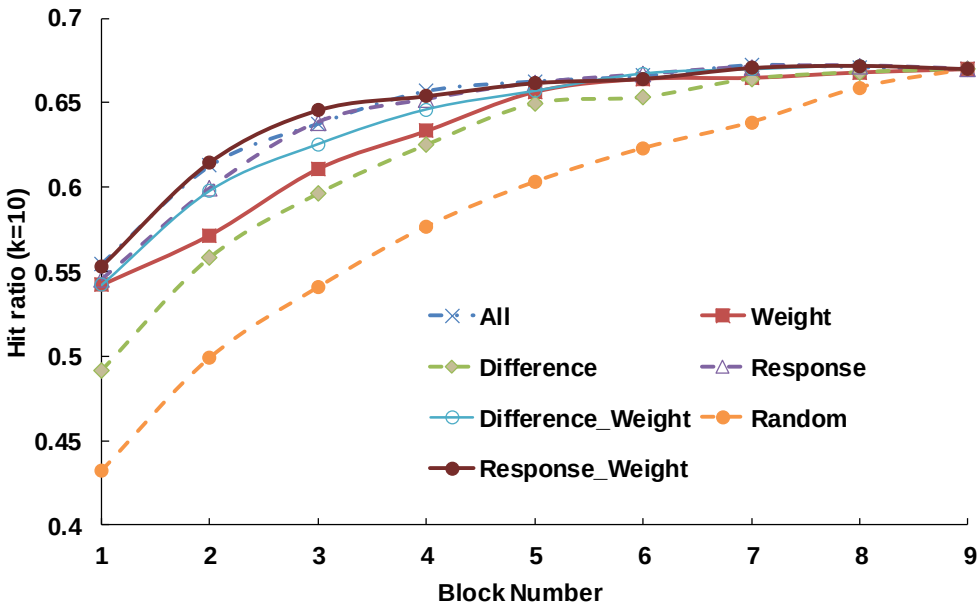


图 3-6 渐进式传输过程中不同传输区块数量以及排序标准的评测

在 3.3.2 节中曾经介绍过，我们使用了视觉哈希码的渐进式传输策略。每个视频关键帧被等分为 9 个相同大小的查询单元，并将查询单元按照其重要性进行排序。之后根据重要性由高到低，渐进式地提取和传输视觉哈希码，并在达到结束条件时，立即结束传输过程。在本小节中，我们评测了渐进式传输的两个重要参数：传输的查询单元个数和查询单元排序指标对移动视频检索准确率的影响。为了比较，我们定义了七种不同的排序方法：（1）Random：随机地将单元进行排序，该方法作为基准方法；（2）Weight：该方法只使用 3.2.2 节中介绍的初始化权重进行排序，如图 3-2 所示，我们将九个单元的初始化权重分别设置为：$\omega_5 = 0.25, \omega_2 = \omega_4 = \omega_6 = \omega_8 = 0.125, \omega_1 = \omega_3 = \omega_7 = \omega_9 = 0.0625$. 这样设置权重是因为视频帧中心内容往往比四周更为重要，并且用户在录

制查询视频时，习惯将视频放在屏幕的中心区域。（3）Response：该方法计算了每个区块的 Hessian 矩阵响应值，并以此作为排序标准。更高的响应值意味着该单元包含了更多稳定的局部特征点，因此需要优先传输。（4）Difference：该方法计算了每个查询单元的相邻帧间颜色直方图的改变，视频内容改变大的单元将优先传输；（5）Response_Weight：该方法将初始权重与 Hessian 响应值相乘，作为排序标准；（6）Difference_Weight：该方法将视频帧差值与初始权重相乘，作为排序标准；（7）All：该方法将响应值、视频帧差值和初始权重相乘，作为排序标准。

为了更加客观地进行评测，我们只使用视觉哈希码进行移动视频检索，并使用命中率作为评测指标。从图 3-6 中展示的结果可以看到，尽管减少了需要传输的特征数量，移动视频检索的命中率并没有显著下降。特别是到传输三个查询单元之后，既可以获得与传输所有查询单元非常近似的命中率。实验结果证明，视觉哈希码渐进式传输的方法可以在不影响查询准确率的情况下，显著降低需要传输的特征规模。除此之外，我们还可以发现，Response、 Response_Weight 和 All 三种排序方法都取得了相似的查询命中率。最后考虑到计算性能和准确率的平衡，我们选择了 Response_Weight 的特征单元排序方法。

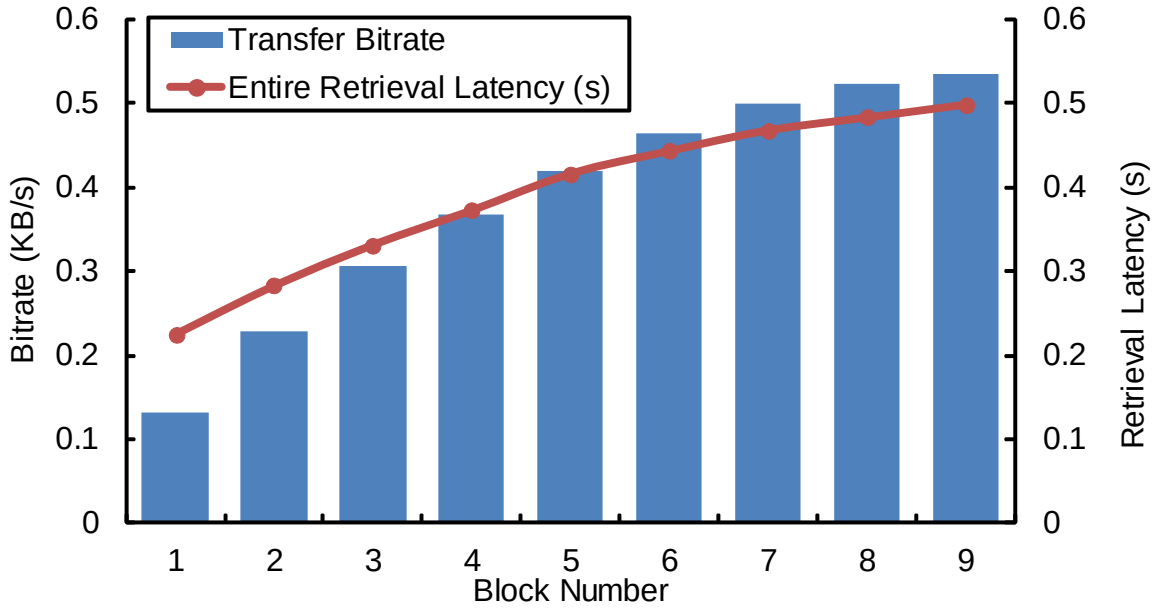


图 3-7 渐进式传输过程中识别比特率和检索延迟的评测

除了对准确率进行验证外，我们还对使用视觉哈希码渐进式传输策略后，移动视频检索总的查询时间的改变进行了评测。图 3-7 展示了视觉哈希码传输比特率（KB/S）和移动视频检索时间延迟随需要传输的单元数目的改变情况。图中 X 轴表示了需要传输的单元的个数，蓝色直方图和左侧 Y 轴表示了视觉哈希码传输比特率，红色趋势线和右侧 Y 轴表示了移动视频检索时间延迟。从图中可以发现，随着需要传输的查询单元的减

少，传输比特率和查询延迟都有了显著降低。如果我们只传输前三个特征单元，则每秒钟只需要传输0.305KB/S的视觉特征，与传输所有特征单元相比，降低了 42.89%的数据传输量。同时，移动视频检索的时间延迟也降低了 33.50%。

3.6.4 移动视频检索性能评测

为了验证移动视频检索的准确率，我们实现并在移动视频检索数据集上比较了以下四种方法：

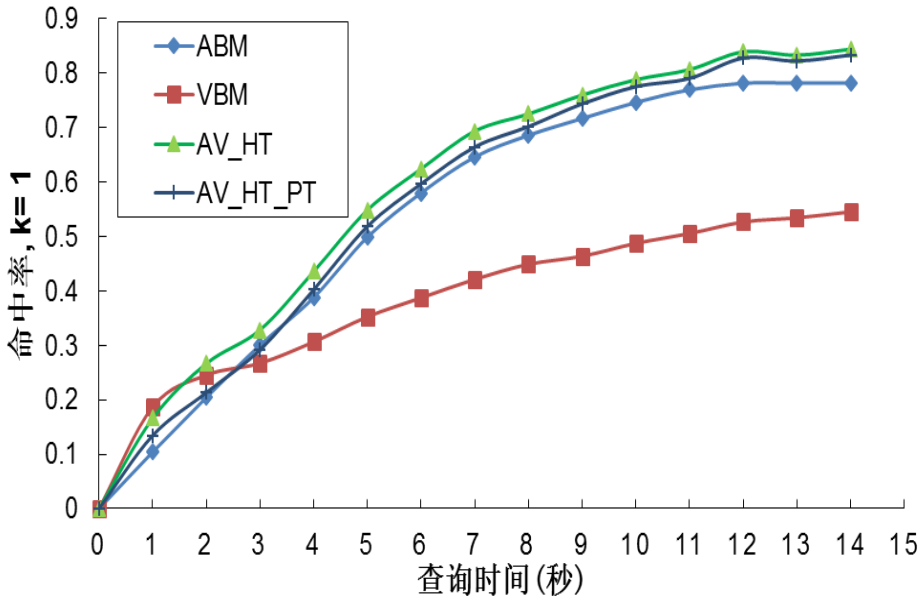
（1）基于音频信息的方法（简称 ABM）。该方法使用音频指纹作为视频签名，使用线性搜索的方法进行匹配。

（2）基于视觉信息的方法（简称 VBM）。该方法使用视觉哈希码作为视频签名，并使用局部敏感哈希的方法进行索引和匹配。

（3）音视频融合的方法（简称 AV_HT）。该方法对音频指纹使用线性搜索，对视觉哈希码使用分层聚类树进行索引，并使用后融合的方法将音视频进行融合。

（4）基于渐进式传输的音视频融合方法（AV_HT_PT）。该方法对音频指纹使用线性搜索，对视觉哈希码使用分层聚类树进行索引，并使用后融合的方法将音视频进行融合。其中视觉哈希码使用了渐进式传输的策略，查询单元使用了 Response_Weight 的排序策略，只传输前三个查询单元。

前两种方法作为基准方法，代表了只使用一种模态信息的视频签名方法。后两种方法使用了后融合的方法对音视频信息进行融合，通过挖掘音视频信息的互补性来提高移动视频检索的准确性。之后，我们分别测试了这四种方法在移动视频检索评测数据集上的命中率，K分别设置为 1，5 和 10。



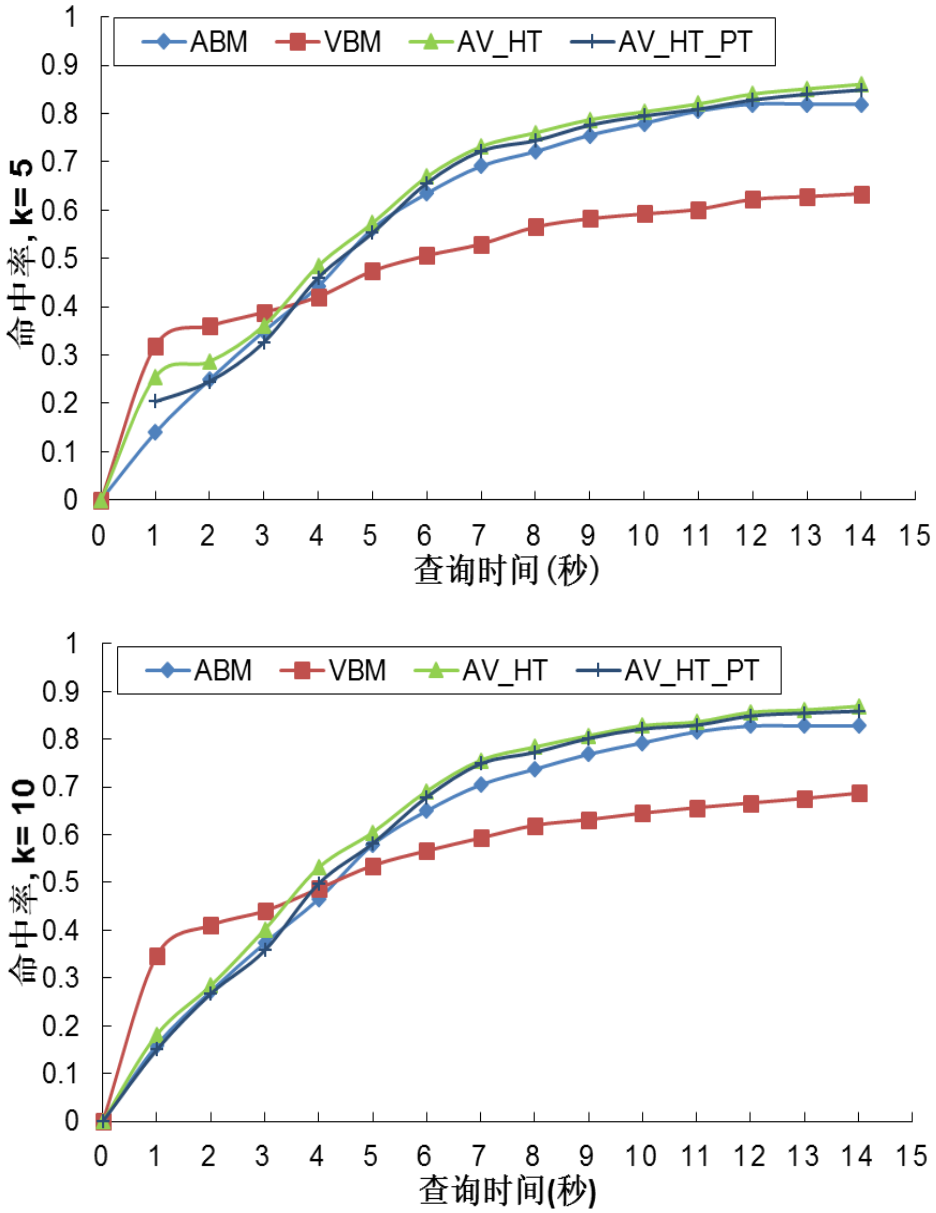


图 3-8 不同视频签名的视频检索性能评测

我们首先在 1,400 个查询视频和 600 小时的视频数据集上测试了 4 种方法的命中率，结果如图 3-8 所示。实验结果证明，与 ABM 和 VBM 相比，音视频融合的方法获得了最高的查询准确率。该结果证明了音频和视频信号的互补性，与单一模态相比，识别能力更强，也使得我们所提出的视频签名对查询视频的变化更加鲁棒。通过比较 AV_HT 和 AV_HT_PT，我们可以发现尽管渐进式传输策略减少了需要传输的视觉特征的规模，但是并没有降低检索的准确率。除此之外，如图 3-8 所示，当查询时间超过 10 秒之后，查询准确率基本不再变化，因此我们可以强制选择 10 秒为最长查询时间。

3.7 小结

由于移动视频检索所面临的特殊挑战，传统视频签名技术已经无法满足移动视频检索的需求，因此，生成和传输具有高识别比特率的视频签名就成了本文研究的第一个重点。为了生成高鲁棒性、轻量级、易传输、易索引的视频签名，我们借助音/视频信息的互补性，分别研究提取了视觉哈希码和基于频谱局部显著性的音频指纹特征。为了提升视觉哈希码的区分能力和压缩率，我们研究比较了不同的二进制编码方法。同时，针对视觉哈希码提取和传输时间复杂度仍然较高的问题，我们研究设计了一种基于 Hessian 显著度加权融合的渐近式视频签名传输方法。最后，经过基于几何信息的一致性验证，进一步提高了视觉签名的鲁棒性。

为了分析移动视频检索的特点，评测本章所提出的视频签名的效果，我们搜集并发布了世界上第一个面向真实环境录制的移动视频检索数据集。在该数据集上的实验表明，本章所提出的融合音/视频的移动视频签名，与使用单一模态的视频签名相比，获得了更高的查询准确率。除此之外，本文所提移动视频签名无论是在提取时间还是在数据规模方面都远小于 CHOG 等经典移动视频签名。同时，本文提出的渐进式传输策略在不降低查询准确率的情况下，大大减少了需要传输的特征规模，降低了查询延迟。经实验验证，该算法每秒只需要传输 0.88KB 的特征，与已有最好方法相比，减少了 33.5%的查询延迟。但是，该章节中使用的视频签名索引和视频匹配方法存在一定的局限性，其查询速度和精度仍然无法满足移动视频检索的精度和速度要求。特别是在音频或者视频中有一个信息不准确的情况下，并不能充分挖掘音视频之间的互补性，造成查询精度下降。因此，在下一章中，我们将讨论如何通过构建新的视频签名索引和匹配方法，充分挖掘音视频融合的互补性，进一步提高移动视频查询的精度和速度。

第4章 音/视频分层哈希索引与匹配技术

4.1 概述

将视频签名压缩为二进制码后，在近似签名的搜索时，可以直接计算签名之间的海明距离，即直接通过位与位之间的异或操作来计算距离，因此匹配速度相比压缩前有了很大提高。尽管如此，对于大规模视频数据，如果只是进行简单的线性扫描，仍然非常费时费力。因此，必须在海明空间中构建索引，提升二进制码的检索速度。文献[56]提出的语义哈希方法直接使用二进制码作为内存地址进行访问，但是当二进制码的位数增多时，很难找到足够的内存用来存放这些二进制码，而且还会造成大量的空桶，浪费内存空间。局部敏感哈希是一种著名的在高维空间中构建索引的方法，也非常适合在海明空间中进行检索[52]。然而，由于它采用随机选择策略，为了实现高精度的检索，必须增加二进制码的长度，以备构建多个哈希表作为候选集，增加检索时命中的概率。为了解决这个问题，文章[53]提出了多探头的方法，通过搜索多个哈希桶的策略来减少哈希表的数量。然而，多桶搜索同样的也会消耗更多的时间。于是，Muja 等人提出了基于分层聚类树的索引方法[65]。虽然分层过滤的策略可以大大减少搜索时间，但是该方法需要构建多棵树来避免查询点的近似点落于探索域的边界外。多棵树结构相应的也增加了存储和查询的成本。除此之外，现有视频签名索引技术只能针对单模态特征进行索引，忽略了多模态信息的互补性。

为了解决以上问题，提升移动视频检索的速度和精度，我们设计了一种音/视频分层哈希索引方法。该索引结构分为两层。第一层是索引的入口，用来过滤掉差异较大的视频签名。该层包含多个独立的索引——一个音频索引和多个视频索引。第一层结构针对音频和视频特征设置了不同的入口，保留了音频特征和视觉特征的独立性。第二层为精确搜索层，由完整的视觉哈希码构成。该层主要考虑不同特征域信息的融合和视频签名的精确查找。我们使用了基于海明距离加权融合的方法，使得整个融合过程既保证了搜索的速度，又能在一个统一的特征空间中动态加权融合不同的模态数据，来获得准确的搜索结果。这种分层结构有两种优势：1. 通过分层过滤策略提高视频签名检索的速度。2. 动态加权融合充分挖掘了不同模态特征间的互补性。

除此之外，与传统的拷贝检测或者近似视频检索不同，移动视频检索中的查询视频只是源视频中任意位置的很小一段视频。因此，移动视频检索是一种典型的子序列匹配任务，目标序列的长度和在源视频中的位置，都是不确定的，需要在目标视频中进行定位和匹配。除此之外，在服务端，每秒钟都会收到新的查询视频关键帧，为了进行实时

检索，服务端也会进行实时的递增式的匹配，并且每秒钟都会将匹配结果实时返回给客户端。这就为移动视频检索中的视频匹配提出了新的挑战。为此，我们提出了一种基于二分图的渐进式视频匹配算法。该方法将匹配上的查询视频关键帧和数据库中的源视频关键帧分别作为二分图中的点，并使用二分图转换和最大二分匹配的方法，进行可变长视频的相似度计算。该方法可以快速的排除大部分不相关的视频，定位到候选视频序列在与源视频中的相似片段。使用二分图来度量查询视频与源视频之间的关系，能够高效的找到最相近的视频序列。而当新的一秒钟内容到达时，只需要在二分图中简单的增加一个点，即可以快速更新二者相似度。而且一旦匹配算法确认获取到了准确的检索结果，就可以自动停止整个检索过程。因此，这种渐进式的视频匹配过程可以显著增加用户的使用体验。下面，我们将详细介绍音/视频分层哈希索引与匹配技术的相关内容。

4.2 音/视频分层哈希索引算法

如图 4-1 所示，音/视频分层哈希索引结构分为两层。第一层是索引的入口，包含多个哈希表：一个音频哈希表（红色）和多个视觉哈希表（蓝色）。音频哈希表是由所有音频签名的二进制码组成的，视觉哈希表是由随机挑选的部分视觉哈希码比特位组成的。第一层中不同的哈希表之间是相互独立的，并且通过索引与第二层相连。第一层的作用是进行粗过滤，过滤掉大部分差异较大的视频。索引的第二层包含一个由完整视觉哈希码组成的索引表。该层的作用是进行精细的视频签名查找和匹配。

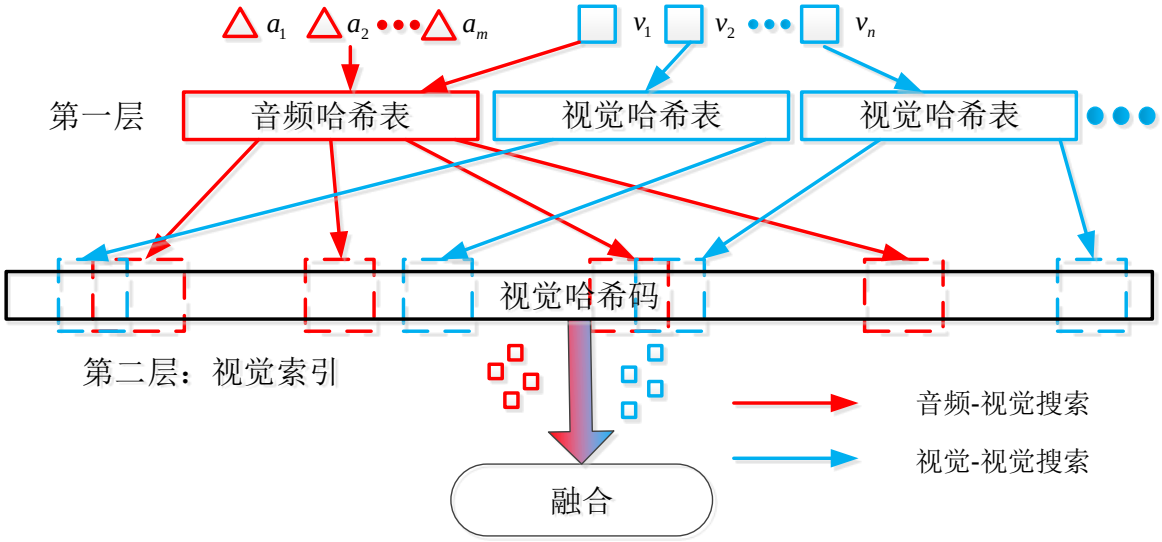


图 4-1 音/视频分层哈希索引的结构图

在移动视频检索的服务端，我们可以将视频签名中的音频指纹特征看作多个音频查询点，视觉哈希码看作多个视频查询点。根据第 3 章所介绍的，用户在拍摄查询视频时，移动端会在后台实时的提取视频签名，并以 1 秒为单位实时传送到服务端。当服务端接

收到 1 秒钟的音频查询点 $A=\{a_1,a_2,\ldots,a_m\}$ 和视频查询点 $V=\{v_1,v_2,\ldots,v_n\}$ 之后，将按照不同的搜索入口，进入不同的搜索过程——音频-视觉搜索和视觉-视觉搜索。两个不同的搜索过程使用了不同的分层过滤策略来检索查询点的近似最近邻。独立的搜索过程可以保持音频和视觉签名的独立性。经过第一层粗过滤之后，算法会继续在第二层中进行精细的查找和匹配，不同搜索过程找到的近似最近邻，会在第二层通过与匹配点之间的海明距离进行加权融合。最后，这些近似查询点再通过几何信息验证进行重排序。这种分层的索引结构主要有两种优势：(1)有效地利用分层过滤的策略来提高视频签名检索和匹配的速度；(2)不同的搜索入口以及基于海明距离加权的融合，既能保证不同模态特征之间的独立性，又能够充分挖掘不同模态特征之间的互补性。下面，本文将详细介绍音频-视觉搜索、视觉-视觉搜索和音频-视频融合的详细过程。

4.2.1 音频-视觉子搜索过程

与视觉特征相比，本文所提取的音频指纹特征更加简洁，每个音频特征点只占用 25 比特位。因此，我们可以直接将音频指纹的哈希码作为内存地址，快速搜索近似最近邻。尽管如此，因为很多不同的视频可能包含相同的音频信息，找到相似的视频指纹后，算法还需要进一步比较他们的视觉相似度。因此，我们使用音频索引作为索引第一层的入口。并通过视频的 ID 和音频的时间偏移值 T，将每个哈希桶与第二层的视觉哈希码联接。通过在第一层的音频索引中过滤掉大部分声音信息不相似的视频，可以显著减少第二层中需要检索的视觉特征点的个数，从而提高特征点查询的速度。在第二层索引中，我们将进一步检索近似的视觉特征点作为候选近似最近邻，与视觉-视觉搜索的结果融合。如图 4-2 所示，整个音频-视觉搜索过程的实现分为如下步骤：

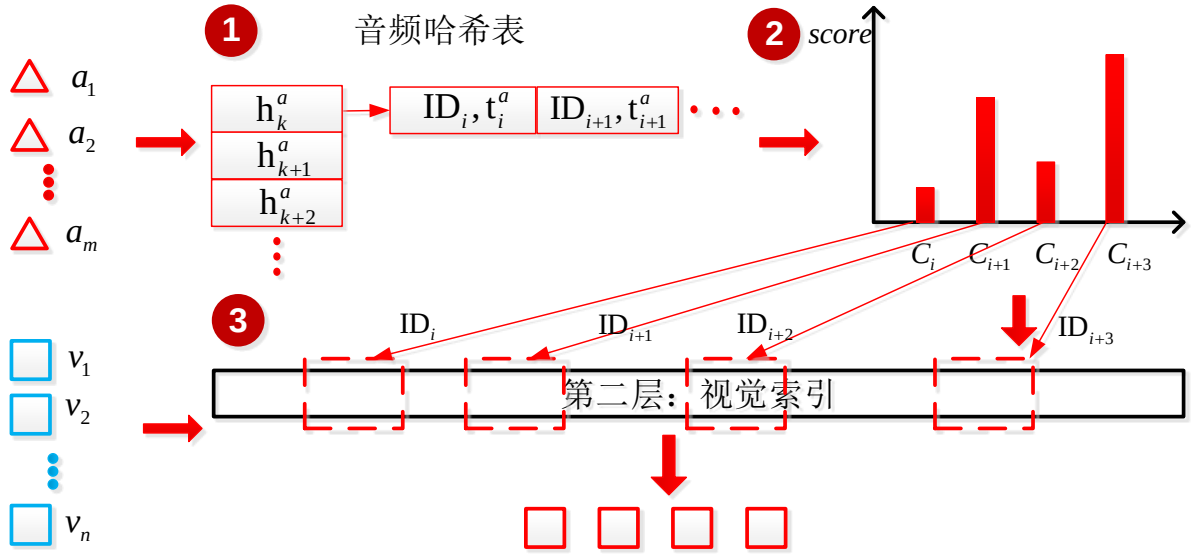


图 4-2 音频-视觉子搜索过程示意图

步骤 1：对于每个音频查询点a_i，算法会在音频索引中自动找到它的前K_a个近似最近邻。得到很多音频匹配对$\{\mathrm{a}_i,\mathrm{t}_i,\mathrm{a}'_j,\mathrm{ID}_j,\mathrm{t}'_j\}$，其中$\mathrm{t}_i$是音频特征点$\mathrm{a}_i$在查询视频中的时间偏移，$\mathrm{a}'_j$是$\mathrm{a}_i$的近似最近邻，$j \le K_a$，$\mathrm{ID}_j$和$\mathrm{t}'_j$是包含音频特征点$\mathrm{a}'_j$的视频ID和在该视频中时间偏移。之后，搜索算法将按照音频匹配对中的ID_j将所有匹配对分配到候选类别$C=\{\mathrm{c}_1,\mathrm{c}_2,\ldots\mathrm{c}_N\}$中去。每个候选类别$c_i$都指代了数据集中的一个源视频。

步骤 2：在该步骤中，搜索算法将会计算每个类别c_i的匹配分数，等于最终分配到该类别的音频匹配对的个数。

1. 我们首先定义$\Delta t = t'_j - t_i$为一个音频匹配对中两个音频特征点的时间差。因为音频特征点在视频中的时间顺序是固定的，所以c_i中的所有匹配对的时间差Δt应该是一致的。因此，我们以此为标准，过滤掉那些错误匹配的音频特征点。
2. 为了达到以上目标，算法首先计算c_i中Δt的直方图，并且找到数字最大的桶。该数值除以匹配对的总数后，就是类别c_i的分数。即查询视频与源视频ID_j的音频相似度。

步骤 3：在此步骤中，搜索算法通过步骤 2 中计算的c_i分数对所有类别C进行重排序，选择前K_v个视频作为候选视频，在第二层索引中继续进行精确匹配。由于候选视频的规模相对于整个数据集已经显著减少，因此，使用线性搜索即可快速找到视觉查询点v_i的前K个近似最近邻。

通过以上搜索步骤可以发现，在音频-视觉搜索中，我们首先使用音频信息找到相似的视频，进而再在这些相似视频中进一步寻找视觉查询点的近似最近邻。最后，对于每个视觉查询点v_i，我们都可以得到K个近似最近邻。这些近似最近邻之后将与视觉-视觉搜索中得到的近似最近邻进行融合。

4.2.2 视觉-视觉子搜索过程

尽管音频-视觉搜索过程可以通过音视频融合的分层过滤法快速找到视觉查询点的近似最近邻，但是如果仅使用音频索引作为第一层入口，当音频信息遭受严重噪声干扰时，将很难在第二层索引中找到准确的视频签名。这里，我们使用多重索引技术，在第一层中构建附加的视觉特征入口，增加视觉-视觉搜索过程来解决这个问题。我们抽取第二层中完整视觉哈希码的部分比特位，在第一层索引中构建m个哈希表，用来作为视觉特征的入口。通过这m个哈希表，我们可以仅通过视觉特征，对视觉特征点进行粗过滤，进而在第二层中进行精确匹配。

如图 4-3 所示，在第二层索引中的视觉哈希码被第一层中h个不同的哈希表索引。每个哈希表中的哈希码h^{sub}是从全部哈希码中随机抽取 40 位得到的。由于缩短了哈希码的长度，我们可以直接在第一层的 m 个哈希表中进行线性搜索。因此，对于一个视觉查询点v_i，我们首先在第一层中的 m 个哈希表中分别找到精确匹配的特征点，再在第

二层中与该桶相连的完整哈希码中进行精确搜索，得到前 K 个近似最近邻。最后，只要将音频-视频搜索过程中得到的近似视频签名和通过视觉-视觉搜索得到的近似视频签名进行融合，即可得到视觉查询点的近似最近邻了。

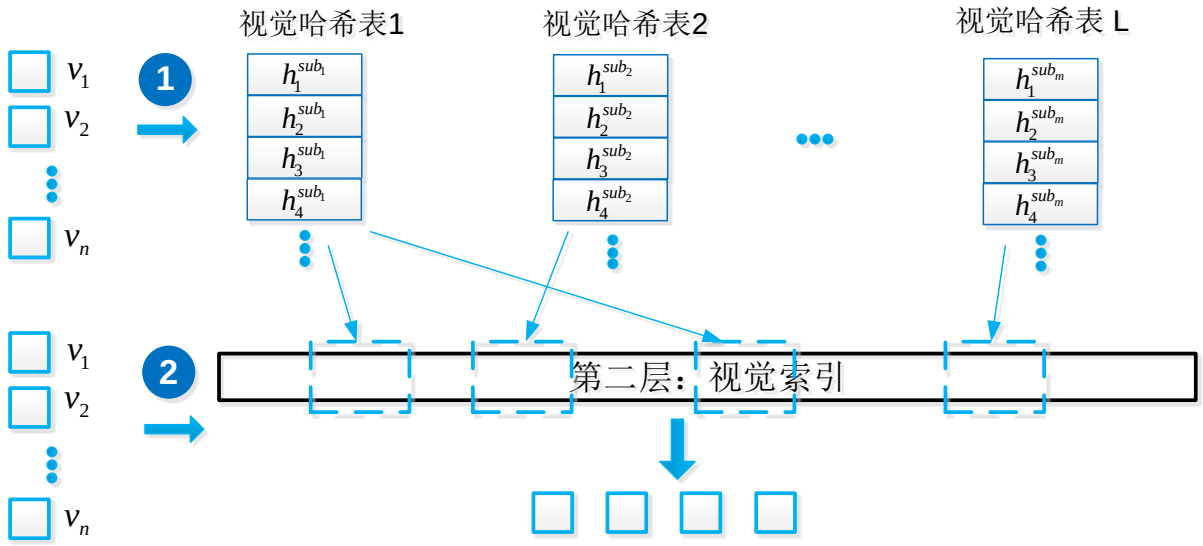


图 4-3 视觉-视觉子搜索过程示意图

4.2.3 音频-视频信息的融合

在音/视频分层哈希索引中，音频信息和视频信息的融合分为两个不同的阶段。第一个融合阶段是在音频-视觉搜索过程中，我们使用更为精简的音频信息进行粗过滤，使用区分性更强的视频信息作为细过滤，可以显著提升搜索的速度。此外，与传统的前融合相比，音频信息和视觉信息的相似度计算是在两个不同阶段进行的，避免了直接拼接两个不同空间向量导致的结构信息的破坏。最后，音频信息的相似度被视觉特征点之间的海明距离所替代。

第二个融合阶段是在第二层，最后两次搜索结果的融合中。对于每个视觉查询点，两个不同的搜索过程都可以得到 K 个不同的近似视觉特征点，并得到查询点与匹配点之间的海明距离。通过该海明距离，我们可以对每个视觉查询点的 $2K$ 个近似视觉特征点重新排序，并选择前 K 个作为最终的搜索结果。与第一个融合阶段严重依赖于音频搜索的准确性不同，第二个融合阶段融合的过程是通过海明距离进行加权融合，自动选择更多参考音频信息或者视觉信息。如果音频信息被严重干扰或者缺失，从音频-视觉搜索中得到的视觉特征点就会排在视觉-视觉搜索的近似特征点之后，被取到最终 K 个近似最近邻的概率就更小，参考的音频信息就越少。反之则越多。因此，第二个融合阶段是根据海明距离，对音视频信息进行动态加权融合。综上所述，两个融合阶段可以更全面的挖掘音频与视觉信息之间的互补性，提供更为鲁棒和准确的移动视频检索结果。

得到最终的 K 个近似视觉特征点之后，我们可以使用霍夫变换的方法找到近似视频关键帧，并通过 3.4.2 节中介绍的几何一致性验证的方法对这些视频帧进行快速的重排

序，得到每秒钟查询视频关键帧的近似视频帧。

4.3 基于二分图的渐进式视频匹配算法

与现有的移动视频检索系统不同，由于移动端的查询信息是以秒为单位连续传送到服务端，本文所提出的移动视频检索的过程也是一个渐进式的匹配过程，即伴随着每秒查询信息的到达，系统可以实时的渐进式的返回查询结果。并且当系统确认得到了正确的查询结果后，就可以自动停止整个搜索过程。渐进式的视频匹配可以显著地减少查询时间，提高用户的使用体验。

在本文中，我们使用二分图的转换和最大匹配算法来实现渐进式匹配的功能。如图 4-4 所示，我们使用一个二分图$G=\{N,E\}$来表示查询视频和源视频。在该二分图中，绿色的点$q_k \in Q$表示第 k 秒到达服务端的查询帧，黄色的点$s_{n,m} \in S$表示源视频V_n中的第 m 个关键帧，符号R_k表示所有与查询q_k近似的视频关键帧。如果$s_{n,m} \in R_k$，就存在一条边$e_{k,m} \in E$. 当下一秒钟查询到达时，我们可以通过算法 4-1 快速的更新整个二分图G_i，然后通过G_i实时的计算查询视频与源视频的相似度。

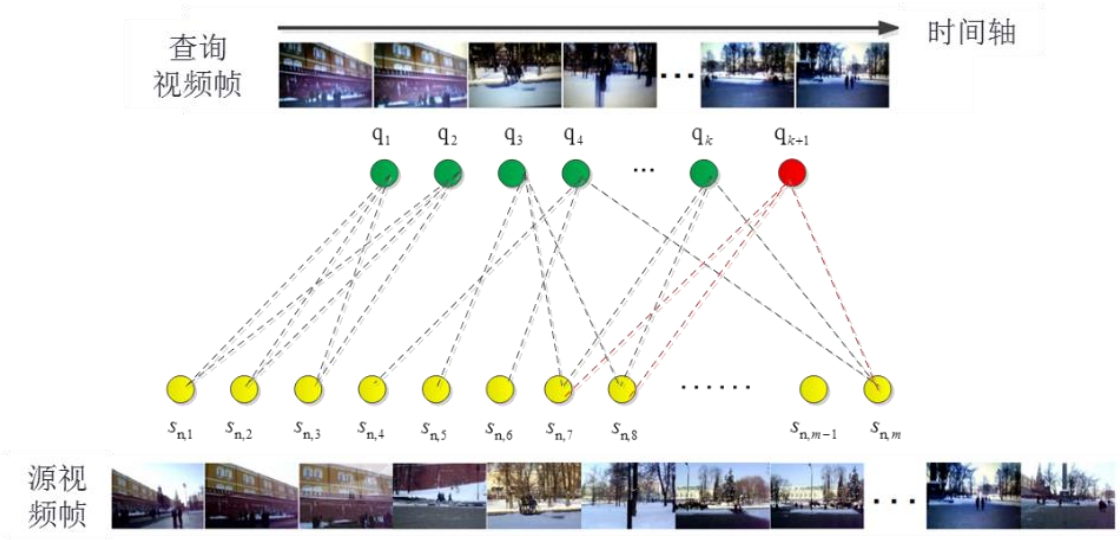


图 4-4 基于二分图的渐进式视频匹配算法示意图

当新的一秒钟查询开始时，算法会在二分图中增加一个新的查询点。之后，图中的边会根据 4.2 节中近似关键帧检索的结果进行更新。对于每个二分图，如果新来的查询帧与源视频中的关键帧相似，则将新增加的点与代表源视频的关键帧的点相连，并更新相应的权值。图 4-4 中红色的点和边显示了更新的过程。如果一个二分图的边没有发生变化，我们就跳过这个二分图。否则，两个视频间的相似度将按照如下步骤进行更新：

表格 4-1 基于二分图的渐进式视频匹配算法

输入：一个新的查询视频关键帧q_{k+1}，q_{k+1}的近似视频帧R_{k+1}
输出：前 K 个近似视频
1: $Q \leftarrow q_{k+1}$
2: $R \leftarrow R_{k+1}$
3: **for** $s_{n,m}$ **in** R_{k+1} **do**
4: 寻找包含$s_{n,m}$的G_i
5: $E_i \leftarrow (q_{k+1}, s_{n,m})$
6: **end for**
7: 调用$W = VideoSimilarScore(G)$
8: **return** R 的前 K 个点
函数 $VideoSimilarScore(G)$
1: **for** G_i **in** G **do**
2: **if** $|E_i|$发生改变**then**
3: 计算G_i的二分图最大匹配M_i
4: **if** $|M_i| > \alpha$ **then**
5: 更新$W_i = Sim(Q, V_i, W_i^a, W_i^v)$
6: **end if**
7: **end if**
8: **end for**
9: **return W**

算法首先计算二分图G_i的最大匹配M_i。如果$|M_i| > \alpha$，则使用公式(4-1)计算查询视频与源视频之间的相似度。

$$\begin{aligned} W_i &= Sim(Q, V_i, W_i^a, W_i^v) \\ &= Sim_a(Q, V_i, W_i^a) + Sim_v(Q, V_i, W_i^v) + Sim_t(Q, V_i) \end{aligned}, \tag{4-1}$$

这里，$Sim_a\left(Q, V_i, W_i^a\right)$代表了音频内容的相似度，可以通过公式(4-2)计算，

$$Sim_a\left(Q, V_i, W_i^a\right) = \frac{\sum w_{k,i}^a}{|Q|}, \tag{4-2}$$

其中$w_{k,i}^a$是查询q_k与视频V_i间的音频相似度，$|Q|$是查询的长度。$Sim_v(Q, V_i, W_i^v)$代表了视觉相似度，可以通过公式(4-3)计算，

$$Sim_v(Q,V_i,W_i^v)=\frac{\sum w_{k,i}^v}{|Q|}, \tag{4-3}$$

其中$w_{k,i}^v$是查询q_k与视频V_i间的视觉相似度。公式$Sim_t(Q,V_i)$代表了查询视频与源视频间的时间相似度。对于二分图G_i的最大匹配M_i，其时间相似度可以通过公式(4-4)获得，

$$Sim_t(Q,V_i)=\frac{LCSS(|Q|,|V|)}{|Q|}, \tag{4-4}$$

其中$LCSS(|Q|,|V|)$是最长公共子序列，可通过公式(4-5)得到。

$$\mathrm{LCSS(i,j)}=\begin{cases}0 & i=0\ or\ j=0\\ LCSS(i-1,j-1)+1 & e_{i,j}>0\\ \max\{LCSS(i-1,j),LCSS(i,j-1)\} & e_{i,j}=0\end{cases}。 \tag{4-5}$$

最后，在计算查询视频与所有源视频的相似性后，系统会自动返回前 K 个最相似的视频。如果查询结果在 3 秒内没有发生变化，则系统会自动结束检索过程。整个匹配算法的算法复杂度为O($|G|\times|N_i|\times|E_i|$)，其中$|G|$是边的数目发生变化的二分图的数量，$|N_i|$和$|E_i|$ 是二分图中点和边的个数。在一次搜索过程中，因为大部分二分图的边并没有改变，因此|G|的数目非常小，算法复杂度也就比较小。

4.4 移动视频检索系统及相关应用

基于前面几节所介绍的移动视频检索关键技术，我们设计了一套完整的移动视频检索系统，使得用户可以自然的通过简单录制一段正在播放的视频，即可进行渐进式的移动视频检索。除此之外，我们还简要介绍了基于该视频检索系统的相关移动应用。

4.4.1 系统框架

图 4-5 展示了我们所提出的移动视频检索系统的整个系统结构。检索系统包含三个部分：移动端，网络传输和服务端。系统结合了第 3 章介绍的移动签名生成与加速算法，本章中介绍的音/视频分层哈希索引技术和基于二分图的渐进式视频匹配算法，提供了实时、准确的移动视频检索，并且显著提高用户的使用体验。

整个系统的操作流程分为离线和在线两个部分。在离线阶段，系统对大规模视频库中的每个源视频抽取融合音/视频信息的视频签名。然后，使用本章中介绍的音/视频分层哈希索引技术将这些视频签名加入到索引中。

在线阶段分为五个检索步骤：1）当用户使用移动客户端录制查询视频时，系统会在后台实时提取轻量级的视频签名。2）提取后的视频签名将以秒为单位渐进式的传送

到服务端。3）当服务端接收到新的一秒钟查询视频签名后，会通过离线阶段构建好的音/视频分层哈希索引进行近似关键帧的搜索。4）得到近似关键帧之后，系统会使用一种快速有效的几何一致性验证方法进行搜索结果的重排序。5）最后，系统使用基于二分图的渐进式视频匹配算法找到最终的近似视频，并实时的将搜索结果返回到移动客户端。如果搜索结果在 3 秒内没有发生任何改变，则系统会自动结束搜索过程，并将最终结果返回给用户。

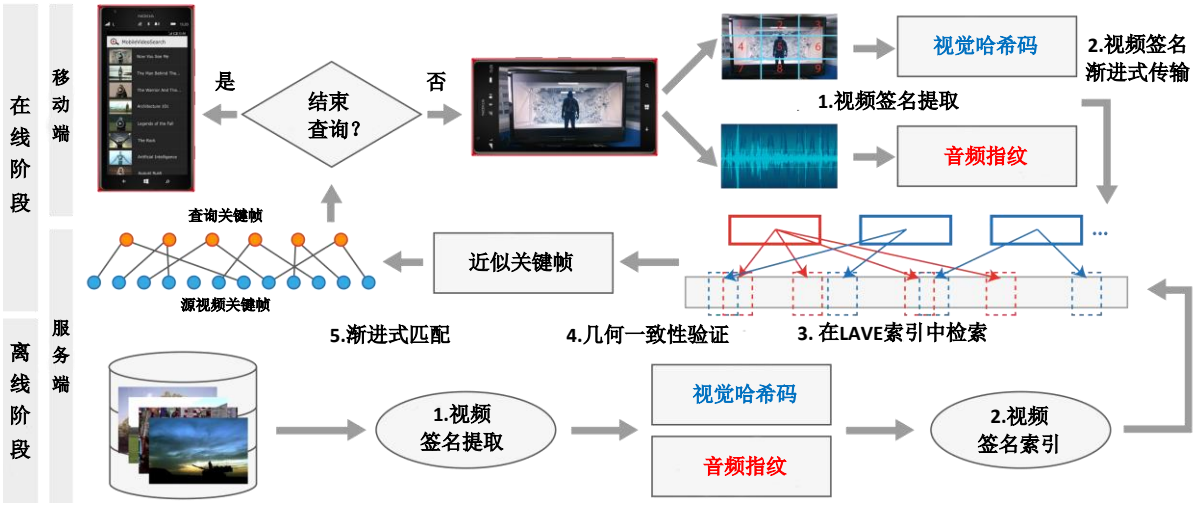


图 4-5 完整的移动视频检索系统流程图

4.4.2 用户交互设计

如图 4-6 和图 4-7 所示，我们在 Windows Phone 8 的设备上设计并部署了移动视频检索系统的用户交互界面。图 4-6 中展示了移动视频检索界面，屏幕中央是视频录制区域，帮助用户查看录制情况。该区域右下角的圆形进度条显示了已经查询的时间。如果超过 10 秒，则系统会自动结束搜索过程。因为搜索过程可能会长达 10 秒钟，如果只是在第 10 秒钟显示搜索结果，会给用户造成很长的等待时间，用户体验较差。因此，我们配合基于二分图的视频匹配算法，提供了渐进式的搜索体验：随着用户拍摄查询视频，相应的查询结果会进行实时更新，并在屏幕右侧进行结果显示。如果用户看到近似的视频，则可通过点击该视频结果立即停止搜索过程。如果发现搜索结果错误，则可以适当调整录制方式，优化查询信息。根据我们的用户主观评测显示，这种渐进式的搜索体验可以显著的提高用户的使用体验。图 4-6 中展示了移动视频检索的结果展示界面，在该界面中，视频缩略图是根据视频片段定位技术，自动选择与用户拍摄的查询视频最相似的视频帧生成的，可以帮助用户快速找到目标视频。用户只要点击结果列表中的相应视频，即可查看视频的详细信息，或者从录制位置开始直接观看视频内容。

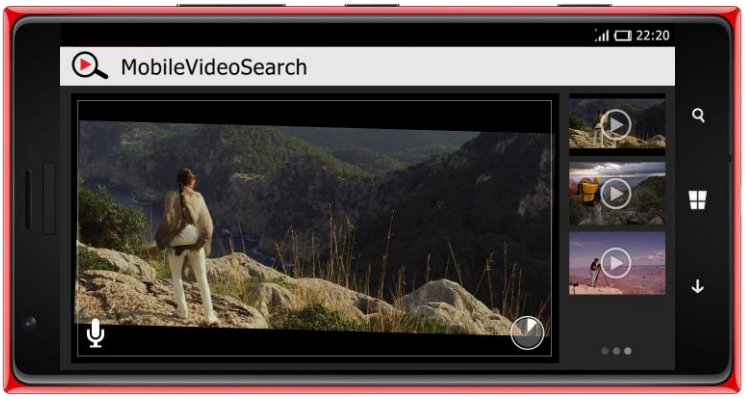


图 4-6 移动视频检索系统的查询界面

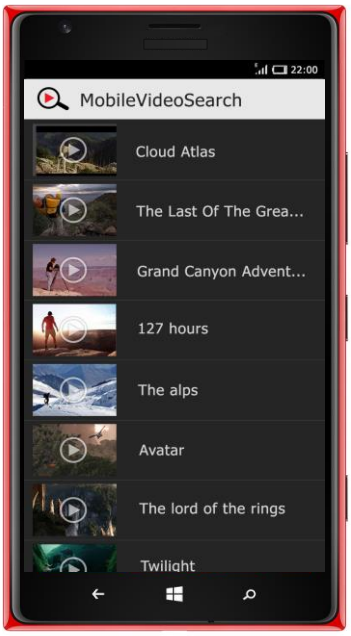


图 4-7 移动视频检索系统结果展示界面

4.4.3 移动视频检索系统的相关应用

4.4.3.1 视频实体检索

作为核心功能，我们的移动视频检索系统能够为移动设备用户提供无处不在的视频实体检索功能。用户只要使用该系统简单拍摄一段感兴趣的视频内容，即可快速获得该视频的近似视频或者相关信息，进而实现更多个性化的功能。比如，如图 4-8 所示，我们可以将系统与国际著名的电影数据网站 IMDB 相连。IMDB 网站包含了世界上最流行、最专业的电影、电视剧内容和权威的评价[10]。当用户在大街上看到一段电影预告片，想了解该电影的详细信息时，即可拿出手机，拍摄一段该预告片的内容，就可获得详细、

[10] “Internet Movie Database (IMDB),” http://www.imdb.com/

专业的资料，比如电影的演员阵容、导演介绍、电影评分、内容简介和观影者的点评等。根据这些信息，用户即可更加容易的决定是否去电影院观看该电影。如果选择是，还可通过移动设备查看打折信息，购买附近推荐影院的电影票等。除此之外，用户还可直接把电影信息分享给其他用户，或者获得相似电影的推荐。

除了视频搜索功能，移动视频检索系统还可以作为视频问答（Video Question-Answering）工具。比如我们可以把系统与新闻网站连接，用户只需要拍摄一段新闻内容，即可找到与该新闻相关的更全面、详细的报道，帮助用户详细了解新闻的来龙去脉。总之，通过视频实体搜索，移动视频检索系统可以作为桥梁，将用户的个人生活与互联网中的群体智慧紧密的连接在一起。

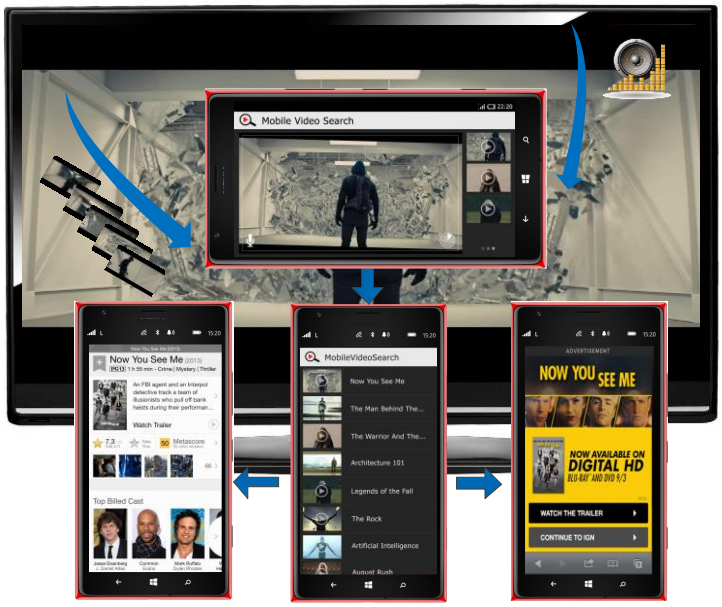


图 4-8 移动视频检索系统的应用实例

4.4.3.2 视频片段定位

除了移动视频检索，本文提出的移动视频检索系统还可以用来定位视频片段。比如，用户在观看一部很长的电影时不得不中途离开时，他就可以使用移动检索系统拍摄几秒钟的视频，即可通过移动互联网上找到该视频，并自动定位到中断的位置。这样，用户可以选择在任何时候任何地点从中断位置继续观看该电影，而无需从开始位置下载观看，节约流量和时间。除此之外，用户还可以使用视频片段定位功能进行动态共享（Living Sharing）。比如用户在电视上看到一段精彩内容，想分享给他的朋友。与原来的使用手机拍摄该片段并上传分享不同，移动视频检索系统可以自动在移动互联网上找到相关视频，并定位到想要分享的片段，直接发送给他的好友。他的好友即可以从他推荐的位置直接观看视频。总之，通过移动视频片段定位，系统可以帮助用户在长视频中自动定位和索引他们感兴趣的内容，方便观看和分享。

4.5 实验结果

本章节将对本文所设计的移动视频检索系统进行全面综合的评测，包括移动视频检索准确性评测、移动视频检索时间评测、移动视频片段定位性能评测和移动视频检索系统易用性主观评测。所有的评测，都在 3.5 节中搜集的面向真实应用环境录制的移动视频检索评测数据集中完成，该数据集包含了1,400个查询视频和19,200个源视频。为了更好的对不同方法进行评测，我们使用了命中率（Hit Ratio）作为评价指标。命中率是指随着查询时间的增加，在前 K 个返回视频中，成功找到目标视频的查询占总查询次数的百分比。由于系统可以返回给用户多个匹配视频，因此，K 分别取 1, 5 和 10。

4.5.1 移动视频检索性能评测

为了验证本文所提移动视频检索关键技术的性能，我们在移动视频检索评测数据集上比较了以下六种方法：

(1) 基于音频信息的方法（ABM）：该方法只使用音频指纹作为视频签名，使用线性搜索的方法进行匹配。

(2) 基于视觉信息的方法（VBM）：该方法只使用视觉哈希码作为视频签名，并使用局部敏感哈希的方法进行索引和匹配。

(3) 基于局部敏感哈希的方法（AV_LH）：该方法对音频指纹使用线性搜索，对视觉哈希码使用局部敏感哈希进行索引，并使用后融合的方法融合音/视频信息。

(4) 基于分层聚类树的方法（AV_HT）：该方法对音频指纹使用线性搜索，对视觉哈希码使用分层聚类树技术进行索引，并使用后融合的方法融合音/视频信息。

(5) 基于音/视频分层哈希索引的方法（LAVE）：使用音/视频分层哈希索引对音频指纹和视觉哈希码进行索引和融合。

(6) 基于渐进式传输的 LAVE 方法（LAVE_PT）：该方法在方法(5)的基础上，对视觉哈希码使用了渐进式传输策略。特征单元排序使用了 Hessian 显著度加权融合的方法，只传输前三个查询单元。

前两种方法作为基准方法，分别代表了只使用一种模态信息的方法。后四种方法都进行了音/视频的融合，不同的是，方法（3）和（4）使用了后融合，而我们提出的方法（5）和（6）使用了音/视频分层哈希索引进行了两次融合。除此之外，方法（3），（4）和我们提出的方法都使用了不同的索引策略。六种方法都使用了基于二分图的渐进式视频匹配算法。下面将详细介绍实验设置、实验结果并进行分析。

4.5.1.1 在完整查询数据集上的评测结果

我们首先在全部 1,400 个查询视频和 600 小时的视频数据集上测试了 6 种方法的命中率，测试结果如图 4-9 所示。从结果中可以发现，与 ABM 和 VBM 相比，我们的方

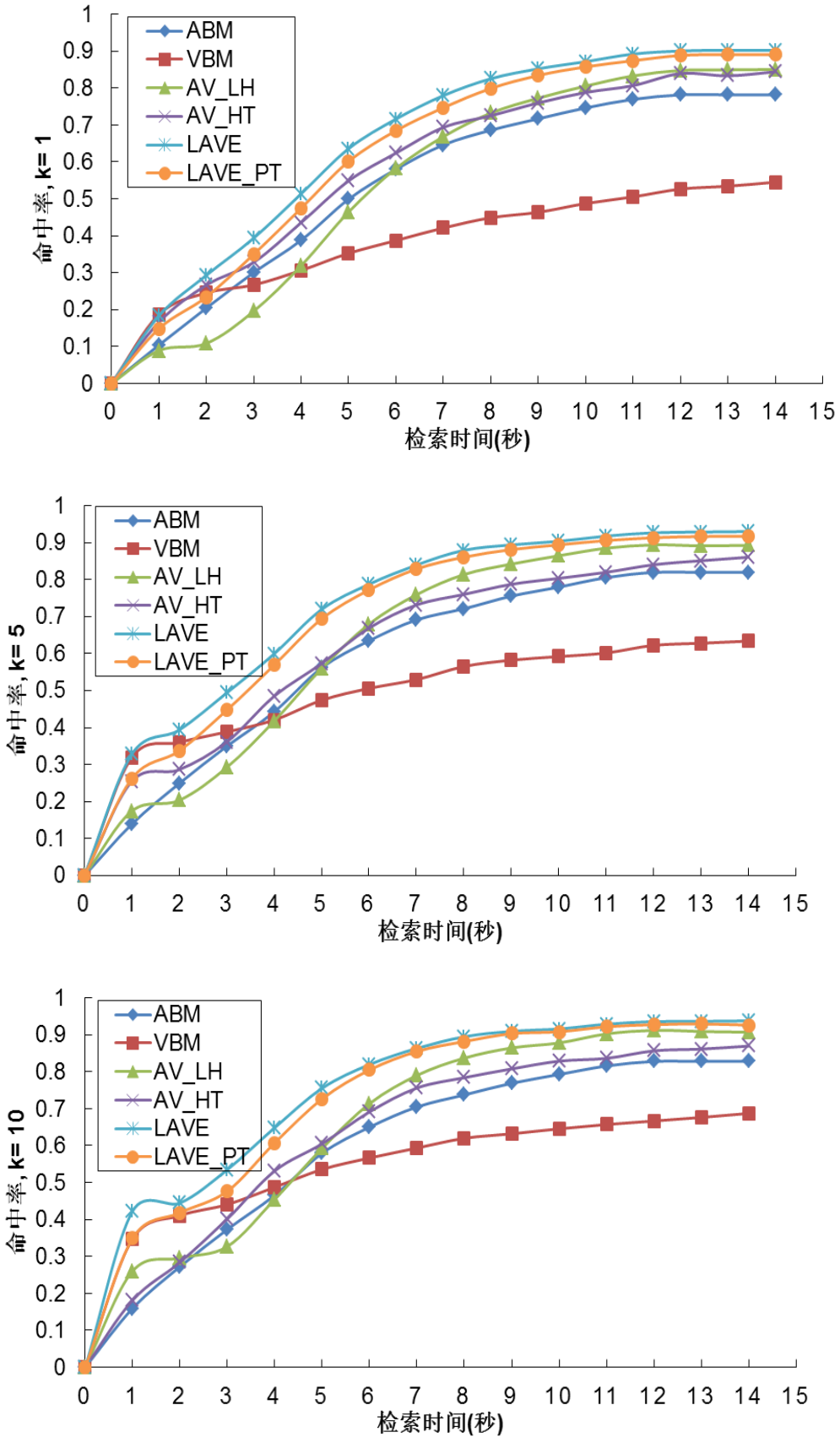


图 4-9 不同移动视频检索方法的检索性能比较，其中(a) k=1, (b) k=5, (c) k=10

法显著提高了查询命中率。该结果证明了与单一模态相比，音视频的融合确实具有更强的识别能力，使其对查询视频的严重图像形变和音频噪声更加鲁棒。在四种都使用了音/视频融合的方法中，本文所提出的音/视频分层哈希索引对音频指纹和视觉哈希码进行索引和融合的方法也取得了最高的命中率。这是因为 AV_LH 和 AV_HT 中使用的后融合方法在融合音/视频信息时使用了平均加权的方法，并没有考虑二者之间的关系。而在音/视频分层哈希索引中，通过分层过滤的方法，我们最终使用了视觉特征之间的海明距离来替代音频特征之间的相似性，因此在最后的音视频融合中，可以动态的考虑音频信息和视频信息的准确性，从而在融合时确定更多考虑音频还是视频信息。该实验结果证明了本章所提出的索引方法能够更有效的挖掘音频和视频信息之间的互补性，获得更好地的融合效果。通过比较 LAVE 和 LAVE_PT，我们同样可以证明，尽管渐进式传输减少了需要传输的视觉特征的规模，却并没有降低检索的准确率。除此之外，由图中结果可以发现，通过有效结合本文所提出的移动签名生成与加速技术、音/视频分层哈希索引技术和基于二分图的渐进式视频匹配技术，我们设计的移动视频检索系统能够在 8 秒钟的查询基础上获得 89.47%以上的准确率，10 秒钟的查询基础上获得 91.59%以上的准确率，比已有最好方法提升了 4%。

4.5.1.2 在不同视频概念集合上的评测结果

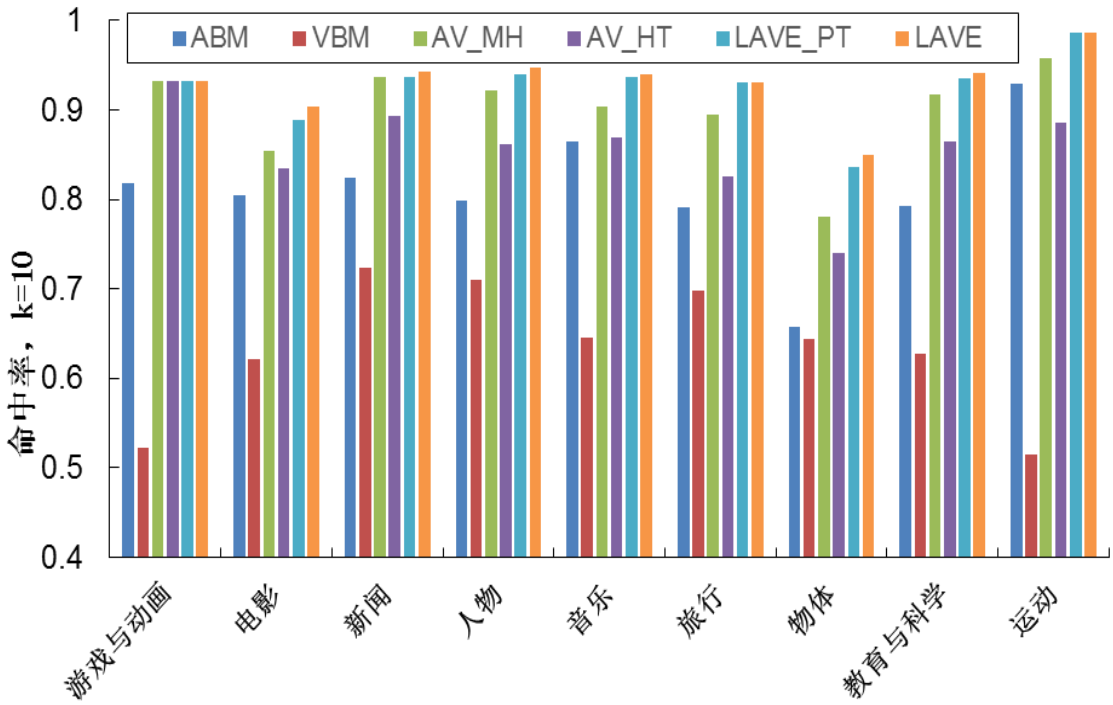


图 4-10 不同移动视频检索方法在不同视频概念集合上的检索性能比较

为了进一步对本文提出的方法进行评测，我们在 9 个不同类别视频上测试了六种方法的命中率。查询时间被设置为 10 秒，命中率指标是从前 10 个返回视频中计算得到的。如图 4-10 所示，在 9 个类别上，LAVE 和 LAVE_PT 方法均取得了比其他方法更高的命中率。除此之外，在“运动”类视频中，AV_LH 和 AV_HT 方法取得了与 ABM 方法近似的命中率，也就是说后融合的方法并没有通过音视频信息的互补提高移动视频检索的性能。相应的，我们最终提出的 LAVE_PT 的方法取得了 6%的提升。通过分析数据，我们发现其原因是很多运动视频包含相似的画面，虽然通过音频信息可以区分这些视频，但是后融合的方法会平均融合视频的音/视频信息的检索结果，反而降低了区分能力。相反的，LAVE_PT 使用的音/视频分层哈希索引可以通过自动赋予音频信息更多的权重来有效的区分这些视频。在“电影”和“人物”类视频上的结果也验证了该假设。

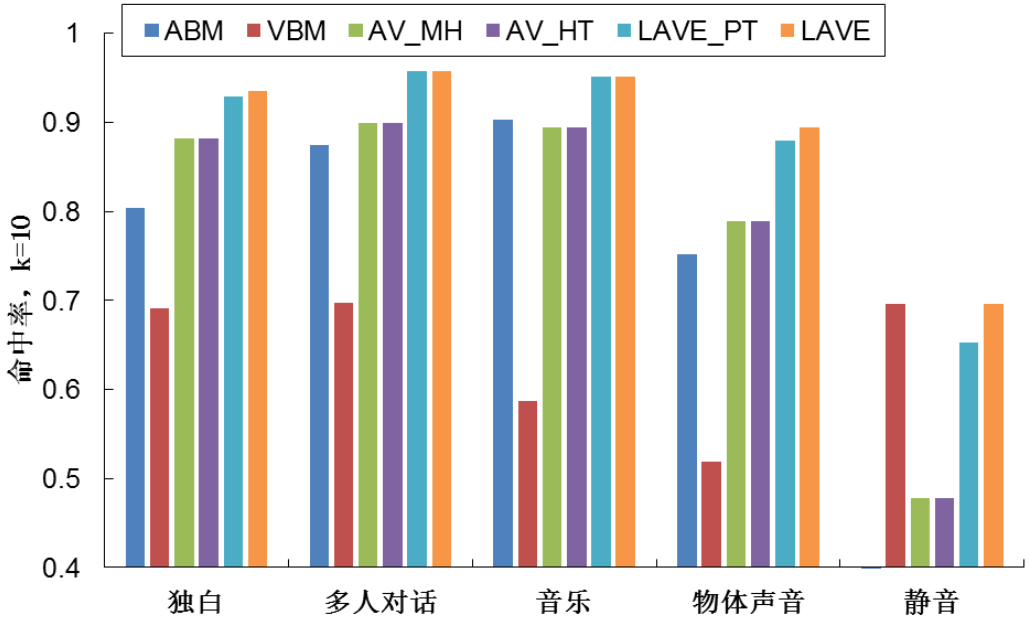


图 4-11 不同移动视频检索方法在不同音频类别视频上的性能比较

4.5.1.3 在不同音频类别上的评测结果

为了进一步验证我们的结论，我们将查询视频根据他们声音信息的不同，分为了以下五种类别：1）独白：在视频中只有一个人说话；2）多人对话：视频中有多个人说话；3）音乐：视频中的声音主要是音乐；4）物体声音：视频中的声音是由动物、交通工具等物体发出的；5）静音：视频中不包含任何声音。从图 4-11 中展示的结果中可以看到，在“多人”和“音乐”类别上的移动视频检索结果，要明显好于其他类别。该结果说明这两种类别的声音区分力更强。除此之外，从结果中还可以发现，我们所提出的两种方法在五种类别中都取得了最好的检索性能。特别是在“静音”类视频这种极端情况，当音频信息根本不起作用时，只依靠视频信息可以获得 69.56%的命中率。在这种情况下，后融合的方法（AV_LH 和 AV_HT）通过平均加权，都拉低了最终的识别准确率。而我

们所提出的音/视频分层哈希索引的方法通过有效地选择参考更多的视频信息，取得了和只依靠视频信息相近的命中率。该实验结果更进一步的证明了我们的结论，与后融合相比，音/视频分层哈希索引的方法可以更全面的挖掘音频与视觉信息之间的互补性，实现更为鲁棒和准确的移动视频检索。

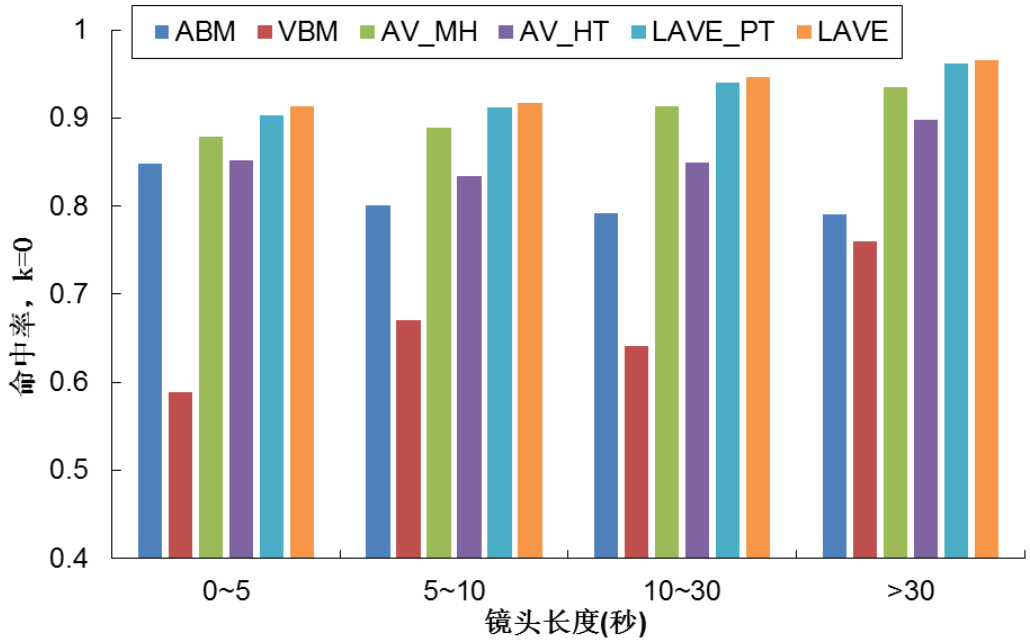


图 4-12 不同移动视频检索方法在不同镜头长度视频上的性能比较

4.5.1.4 在不同镜头长度上的评测结果

除了测试了在不同音频类别视频上的检索性能，我们还将视频按照镜头长度进行分类，测试了 6 种方法在不同视觉类型视频上的检索性能。通过视频的平均镜头长度，我们将视频分为平均镜头长度在 0~5 秒，5~10 秒，10~30 秒和大于 30 秒四种不同类型的视频。平均镜头长度的不同，代表了视频所描述视觉内容的变化程度。6 种不同的移动视频检索方法在这四种不同类型视频上的检索结果如图 4-12 所示。从图中可以发现，与其他四种方法相比，本文所提出的方法都获得了最好的查询性能。更重要的是，从结果展示中还可以发现，当平均镜头长度越长时，检索命中率越高。这是因为镜头越长，视频中包含的冗余信息越多，视频内容越稳定，对于视频形变的适应性也越高，查询性能也就越好。因此，对于镜头长度较短，镜头切换频繁的视频，我们可以考虑抽取更多的视频帧，以此来提高移动视频检索的性能。

4.5.2 移动视频检索时间评测

对于即时视频检索来说，检索时间是重要的评测指标。因此，本文分别测试了所提出的移动视频检索系统在服务端视频检索时间、结果传输时间和总体查询时间。

4.5.2.1 服务端视频检索时间

为了测试服务端的视频检索时间，我们在一台搭载了 Intel(R) Xeon(R) 3.33 GHz 主频的处理器，48 GB 内存的服务器上部署了移动视频检索系统的服务端。通过在 600 小时规模的视频库上进行 1,400 个视频的查询，我们得到平均查询时间为 127ms，其中包括 119ms 在音/视频分层式哈希索引中的查询时间和 8ms 的渐进式视频匹配时间。作为比较，我们还实现了局部敏感哈希索引和基于分层聚类树的索引方法，分别需要 1453ms 和 593ms。该结果证明了本文所提出的音/视频分层哈希索引与基于二分图的视频匹配技术出色的视频检索性能。

4.5.2.2 结果传输时间

作为一个完整的移动视频检索系统，我们还需要加入查询结果回传的时间。因为我们只需要传输视频的缩略图和标题，如果传输前五个查询结果，则每秒钟需要传输 120Kb 的数据。具体的时间延迟依赖于测试时的网络环境。理论上，对于一个典型的带宽为 320Kb/s 的 3G 网络，传输所有检索结果需要大约 375ms。

4.5.2.3 系统整体检索时间

在本文设计的移动视频检索系统中，以上 4 个组件是可以并行运行的。比如音频和视频特征可以同时提取，服务端接收到部分特征点后可以立即进行查询等。为了测试整个系统在真实网络环境下的查询时间，我们使用设备 I 在真实的 WIFI 网络和 3G 网络下分别检索了 100 个不同的视频。最后，查询系统的整体时间在 WIFI 网络下为 626ms，3G 网络下为 781ms。由此可见，本文所提出的系统可以进行实时的移动视频检索。

4.5.3 移动视频片段定位性能评测

在 4.4.3 节中，我们曾经提到本文所提出的移动视频检索系统的一个重要应用是进行视频片段定位。因此，我们在移动视频检索数据集上测试了使用 LAVE_PT 方法进行视频片段定位的准确性，并与 ABM 和 VBM 方法进行了比较。为了有效比较这三种方法定位的准确性，我们实用了定位偏移量作为评测指标。定位偏移量 $\Delta l = \left| l_t - l_g \right|$，其中 l_t 是通过程序找到的视频片段在源视频中的开始时间；l_g 是由视频录制者标注的准确的开始时间。查询的视频长度被设置为 10 秒钟，当计算定位准确率时，我们只考虑准确命中源视频的情况。视频定位的结果如图 4-13 所示。其中横坐标表示视频片段定位偏移区间，直方图和左侧纵坐标表示定位偏移值落在该区间的查询次数，线段和右侧纵坐标表示左侧直方图的累计值。如图中结果所示，与 ABM 和 VBM 方法相比，我们提出的方法获得了最高的视频片段定位准确率。该实验结果进一步证明了音视频融合的重要性。根据统计结果显示，有 78.72%的查询视频的定位偏移量小于 4 秒。而因为我们数据集中有 90.91%的视频镜头长度大于 4 秒，我们可以认为这些视频定位都落在了同一镜头

内。由此可见，视频片段定位的结果基本满足应用需求。

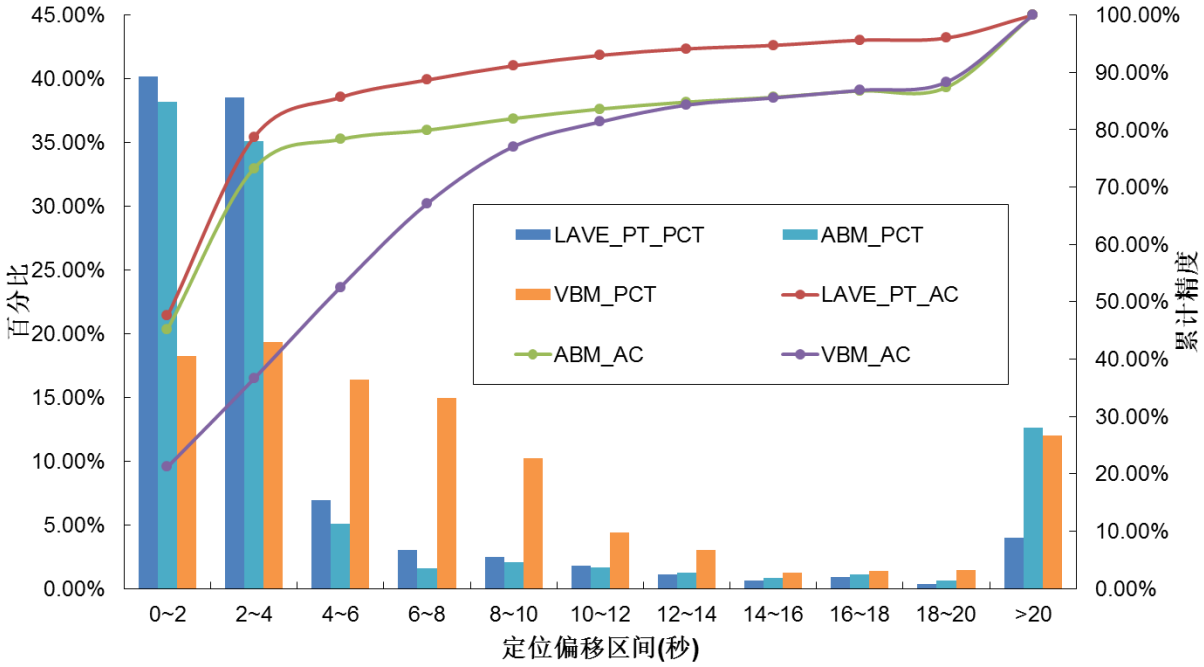


图 4-13 移动视频片段定位性能评测

4.5.4 移动视频检索系统的易用性主观评测

为了验证我们设计的移动视频检索系统的易用性，我们邀请了 12 名志愿者前来体验和使用我们的 LAVES 系统和目前流行的移动视频检索系统 IntoNow 和 VideoSurf，并进行易用性主观评测。被邀请的自愿者包含 3 名女性和 9 名为男性。他们都是在校研究生或者公司职员，年龄从 22 周岁到 36 周岁不等。因此他们都已经参加了之前面向真实环境的移动视频检索数据集的查询视频录制，所以经过大约 3 分钟的指导和演示之后，所有志愿者都熟悉了本文所设计的移动视频检索系统的具体操作。通过调研，我们发现有 10 名志愿者认为该系统提出的渐进式搜索过程体验感非常好，使得视频搜索过程变得更有趣。特别值得一提的是，当志愿者得知我们的系统并不限制用户只使用音频或者视频信息时，所有人都认为这样非常方便，并且将要尝试只使用一种信息来进行视频检索。

在志愿者熟悉使用所有系统之后，我们要求他们使用该手机应用完成两个任务，并对我们的系统做出主观评价。这两个任务分别是：

任务 1：每个志愿者在 19,000 个源视频中挑选 5~10 个视频，尝试使用 LAVES 系统去搜索这些视频的源视频。因为我们的系统还没有公开发布，因此目前只能在实验室环境下测试，所有的源视频都来自于本文搜集的移动视频检索数据集。志愿者可以选择只使用音频、视频或者二者都用。

任务 2：在本任务中，我们要求志愿者比较本文提出的 LAVES 移动视频检索系统和目前已有的其他移动视频检索系统：IntoNow 和 VideoSurf。在熟悉这三种系统后，志愿

者被要求自由使用这三个系统 20 分钟，然后回答我们准备的调查问卷，对三种系统的可用性、易用性和用户体验进行评测。

表格 4-2 不同移动视频检索系统易用性主观评测结果

ID	问题	IntoNow	VideoSurf	LAVES
1	系统吸引力	3.4	3.8	4.2
2	搜索自然度	3.8	3.6	4.3
3	系统实用性	3.3	3.2	4.1
4	系统易用性	4.3	4.2	4.7
5	系统效率	3.1	3	4.3
6	系统准确性	2.8	3	4.3
7	综合推荐指数	3	3.2	4.2
8	渐进式搜索体验	---	---	4.4

注：采取 5 分制，5 分为满分

在任务 1 中，所有志愿者总共搜索了 120 个视频，其中 104 次都成功找到了目标视频。除此之外，所有人都认为渐进式搜索过程减少了他们的搜索时间，提高了他们的搜索体验。根据我们的记录，所有 120 次查询的平均时间为 8.5 秒。对于任务 2，志愿者使用 IntoNow、VideoSurf 和 LAVES 系统之后的意见反馈如表格 4-2 所示。该反馈结果显示了 LAVES 系统相较于其他两个移动视频检索系统的优势：

1） 使用渐进式传播的轻量级音视频签名可以显著降低查询时间；

2） 融合音/视频的分层式哈希索引可以有效的索引大规模视频，提升视频检索的速度和精度；

3） 所有用户都对不限制使用音频或视频信息作为查询感到满意；

4） 渐进式搜索体验可以显著提高用户对程序的满意度。91.67%的志愿者都对渐进式搜索过程感到满意，认为这使得移动视频检索更加自然。他们对渐进式搜索体验给出了 4.4 分的评价。

因为以上这些优势，所有志愿者都表示愿意在他们的移动设备上安装该程序并且推荐给其他人使用。除此之外，志愿者还给出了很多有价值的建议，比如“源视频库中增加更多流行的视频”、“增加搜索结果的分享功能”等等。

4.6 小结

在本章中，我们主要对移动视频检索技术中的索引和视频匹配关键技术进行了研究。目前已有的二进制哈希索引算法都是针对单一类型的特征进行处理，而使用单一特征进行视觉信息查询的性能是有限的。随着不同模态信息的丰富，研究从多种特征模态挖掘

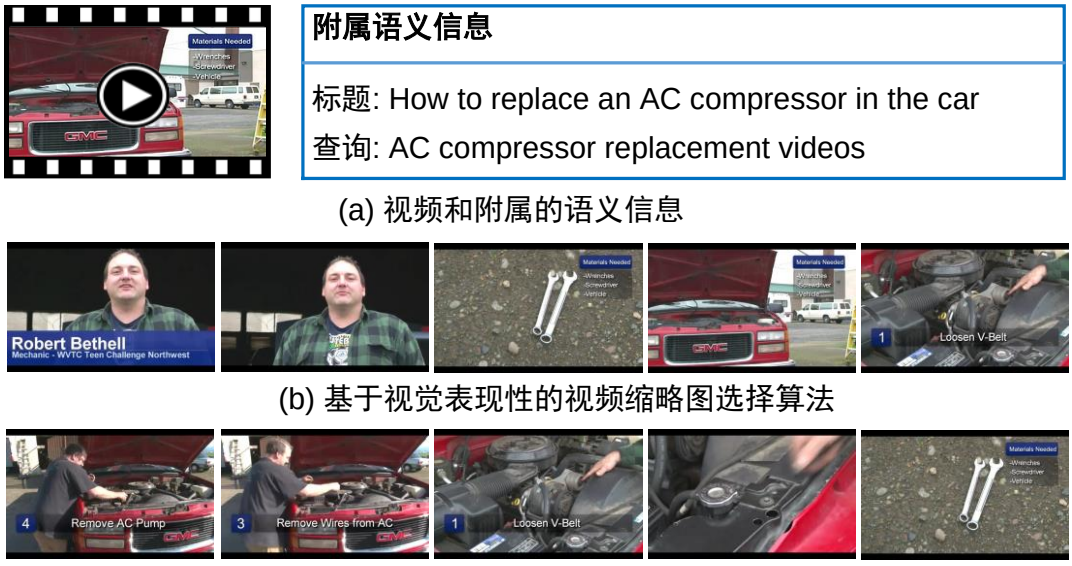
特征之间的关联，实现多模态移动视频检索成为一个重要的问题。而我们提出的音/视频分层哈希索引技术，借助音视频信息的融合，在提高搜索速度的基础上，实现了在统一的特征空间中融合不同的模态数据来获得准确的搜索结果。实验证明，我们提出的索引技术具有如下优势：1. 与现有二进制哈希索引技术相比，通过音/视频分层过滤策略显著提高了视频签名检索的速度；2. 与常用的前融合和后融合技术相比，能够通过两次融合，充分挖掘不同模态特征间的互补性，提高了视频签名的识别能力和鲁棒性，并显著提高了移动视频检索的精度。同时，本文提出的基于二分图的渐进式视频匹配算法，充分考虑了移动视频检索子序列匹配任务中目标序列的长度和在源视频中的位置都是不确定的特点，利用二分图转换和最大匹配技术对查询信息和源视频内容进行渐进式的匹配，定位候选视频序列，实时更新查询结果并返回给用户。根据检索性能客观评测和易用性主观评测结果证明，该匹配技术在提高查询准确率的基础上，显著提高了用户的使用体验。

除此之外，我们结合第 3 章和第 4 章中提出的移动视频签名生成与加速技术、音/视频分层哈希索引技术和基于图模型的渐进式视频匹配技术，设计并实现了一套完整的移动视频检索系统，并提出了相关的应用方案。最后，在本文所构建的面向真实环境的移动视频检索评测数据集中，我们对该移动视频系统进行全面、综合的评测。与现有最好的视频索引与匹配技术，以及流行的移动视频检索系统相比，本文方法在移动视频检索查准率、检索速度、视频片段定位性能以及易用性方面都取得了最好的效果。

第5章 基于视觉-语义深度嵌入的视频缩略图选择

5.1 概述

视频缩略图（Video Thumbnail）的主要作用是展现视频的核心内容，帮助人们快速了解视频信息。特别是在移动视频检索中，由于网络带宽和流量的限制，用户无法通过快速浏览视频来获取其主要内容，视频缩略图就成了帮助用户了解视频内容，确定目标视频的最有效方式。但是，传统视频缩略图选择算法主要选择具有视觉代表性的关键帧作为视频缩略图[78][83][79]，忽略了视频的附加语义信息（如视频的题目、描述、查询等）。如图 5-1 所示，一个标题为“汽车空调压缩机更换”的视频，其视频缩略图不仅要代表视频的视觉内容，还应该和该标题所展现的语义信息有关。特别是当移动用户给出了搜索“汽车空调压缩机更换视频”的查询时，如果返回结果中的缩略图和查询相关，可以帮助移动用户快速找到目标视频，提高移动视频检索的效率和用户体验。



(a) 视频和附属的语义信息

(b) 基于视觉表现性的视频缩略图选择算法

(c) 基于深度视觉-语义嵌入模型的视频缩略图选择算法

图 5-1 查询相关的视频缩略图选择方法

在移动视频检索中，如果用户选择拍摄一段视频作为查询进行视频检索，我们可以通过 4.4.3.2 节中提出的视频片段点位技术，为用户找到最佳的视频缩略图。但是当用户选择通过文本信息进行移动视频检索时，则需要返回与查询文本相关的视频缩略图。因此，本章的主要研究内容就是通过有效结合视频的附加语义信息，特别是用户的查询语句，来生成查询相关的视频缩略图。如 2.5 节中介绍，查询相关的视频缩略图选择存在

的关键问题是怎样快速、有效的计算用户文本查询与视频缩略图之间的相似度。为了解决该问题，我们提出了一种基于多任务学习（Multi-Task Learning）的视觉-语义深度嵌入（Deep Visual-Semantic Embedding）技术。该技术可以有效地将多样的视频附加语义信息与视频视觉内容连接在一起，并应用在移动视频检索中的查询相关的视频缩略图选择算法中。其主要思想是通过视觉-语义深度嵌入模型，将文本和视觉信息映射到隐含的语义空间。这样两种不同空间的信息就可以通过它们在语义空间的投影向量直接计算相似度。除此之外，为了充分挖掘视觉与语义的关联性，得到具有普适性的嵌入模型，我们使用了大规模带有用户点击信息（Click-through-based）的图像和视频数据集，并通过多任务深度学习的策略来减少图像和视频之间的语义鸿沟。

训练得到视觉-语义深度嵌入模型之后，使用该嵌入模型进行查询相关的视频缩略图选择的具体步骤如下：首先，对视频抽取八种不同的视觉表现属性，并利用这些属性选择了 20 帧最具视觉表现性的视频关键帧作为候选缩略图。之后，利用嵌入模型将用户的查询语句和候选缩略图映射到隐含的语义空间，并计算二者之间的相似度。最后，将每个候选关键帧的视觉表现性和查询相关性融合，来挑选最终返回给用户的缩略图。如图 5-1 所示，与(b)中的传统视频缩略图选择方法相比，本文所提方法选择的视频缩略图(c)不仅完整的表现了视频的视觉内容，还能有效帮助移动视频检索用户快速找到目标视频。下面，我们将详细介绍基于多任务学习的视觉-语义深度嵌入技术和查询相关的视频缩略图选择算法的具体内容。

5.2 视觉-语义深度嵌入技术

视觉-语义深度嵌入理论是首先由文献[93]提出的。该技术利用在文本域中学习到的语义知识，迁移到视觉语义识别模型中去。通过将图像和其文本标签嵌入到隐含的语义空间，该技术可以通过语义向量直接计算两者之间的相似性。更重要的是，对于在训练集中没有出现过的标签，视觉-语义嵌入技术仍然能够通过标签在隐藏语义空间的表征找到相似的图像。因此，尽管我们的训练数据集不可能完全覆盖文本查询空间，仍然可以通过视觉-语义嵌入技术来预测任何可能的查询与视频缩略图之间的相似性。

视觉-语义深度嵌入模型的结构如图 5-2 所示。对于输入的文本查询，该模型使用了文本嵌入模型将其映射到隐含的语义空间。文本嵌入模型是自然语言处理领域中常用来将文本信息转换为带有丰富语义信息稠密向量表示的工具。在这里，我们使用了由美国斯坦福大学 Access Forbidden 等人提出的“GloVe”文本嵌入模型[11]。该模型在包含 8,400 亿单词的语料资料上训练得到。我们之所以选择“GloVe”，是因为它在单词类比（Word

[11] “GloVe,” http://nlp.stanford.edu/projects/glove/.

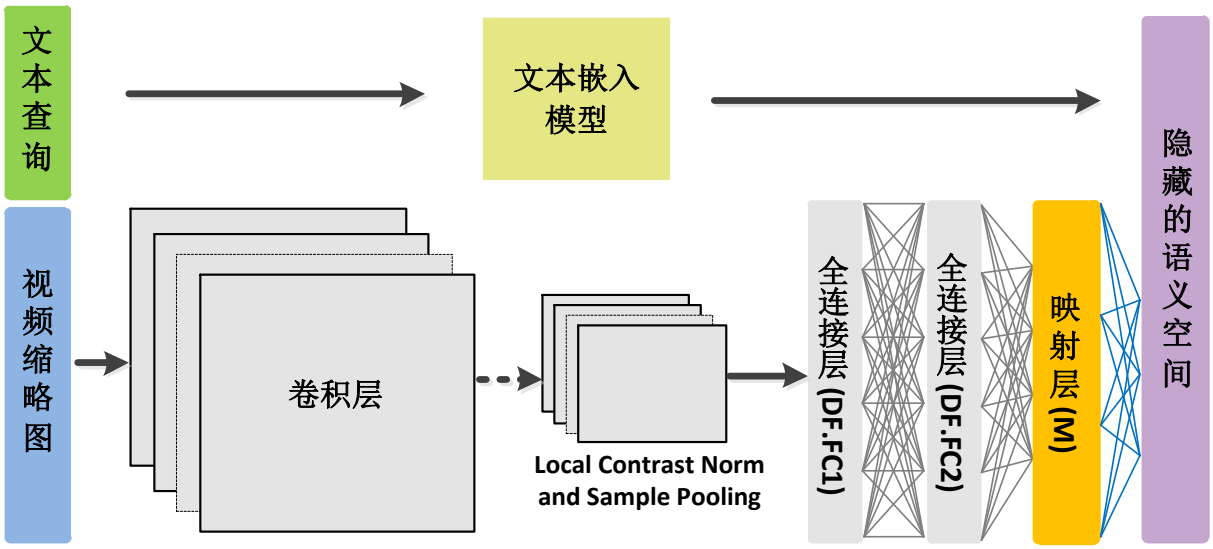


图 5-2 视觉-语义深度嵌入模型的结构图

Analogy）、单词相似度分析（Word Similarity）等任务中的出色表现[101]。尽管如此，本文提出的算法并不依赖于某一种特定的文本嵌入模型，“GloVe”模型将来可以被其他更有效的文本嵌入模型替换。在具体实现时，我们按照文献[93]中的建议，为了获得训练的速度、语义表达的质量和最终模型性能之间的平衡，选择将文本查询映射为 300 维的语义特征向量。

对于输入的视频缩略图，我们首先使用了深度卷积神经网络（Deep Convolutional Neural Network，简称 CNN）将其映射到隐含的语义空间。具体的网络结构是由多伦多大学的 Alex Krizhevsky 提供的“Cuda-convnet”程序实现的[12]。该网络模型是在“ILSVRC-2012”数据集上训练得到的，并且获得了 1000 类分类问题的第一名[102]。原始的 CNN 网络包含两部分：1）输入层，五个卷积层（Convolution Layers）和最大池化层（Max-pooling Layers）；2）两个全连接层（Fully Connection Layers）“FC1”、“FC2”，和最终输出图像在 1000 个类别上概率分布的输出层（Output Layers）。这里，为了将视觉信息映射到隐含的语义空间，我们将输出层的 1000 个类别替换为与视频图像相关联的文本查询的语义特征向量。同时，将原本的输出层替换为映射层 M。然后，原有模型的损失函数也被替换为公式(5-1)的形式，

$$loss(v,\vec{t}^{+},\vec{t}^{-})=\sum \max[0,\gamma-\vec{t}^{+}M\vec{v}+\vec{t}^{-}M\vec{v}], \tag{5-1}$$

其中，$\vec{t}^{+}$ 是查询的语义特征向量，$\vec{v}$ 是与查询相关的视频缩略图在深度神经网络全连接层 FC2 的输出，$\vec{t}^{-}$ 是与该缩略图无关的其他文本查询的语义特征向量，M 是映射层 M 中训练得到的参数矩阵，γ 是边缘参数（本文中设为 0.1）。该损失函数是视觉-语义相

[12] “Cuda-convnet,” https://code.google.com/p/cuda-convnet/.

似性（Dot-product Similarity）和铰链秩损失（Hinge Rank Loss）的组合。该损失函数的目标是将图像视觉信息映射到隐含的语义空间后，得到的语义特征向量与相关联查询的语义特征向量的相似度要大于与无关联文本的相似度。

因为参数矩阵 M 被设置为 CNN 网络中的映射层，我们可以使用随机梯度下降法对整个视觉-语义嵌入模型进行训练。在训练过程中，我们选择从视频搜索网站 Bing 的查询日志中搜集到的带有用户点击信息（Click-through）的视频数据集作为训练集。该数据集包含了 50 万个用户查询和点击记录，以三元组 $\{query, URL, click\}$ 的形式存放，其中 $query$ 是用户输入的文本查询，URL 是用户点击的视频的超链接，$click$ 是用户的点击次数。这里点击次数越多，说明用户输入的文本查询与被点击视频之间的关联性越高。因此，为了保证二者之间确实存在关联，我们过滤掉了点击次数小于 5 次的点击集合。对于过滤后的三元组集合，我们通过 URL 下载了视频的标题和缩略图，并使用自然语言处理工具 NLTK[103]删除 $query$ 和视频标题中的非英语单词，停用词和单词后缀[13]。然后，我们将视频缩略图作为 v，$query$ 和视频标题中的单词作为 $\vec{t}^{+}$，随机挑选其他单词作为 $\vec{t}^{-}$。经过以上处理，我们最终得到了 41 万个三元组 $\{v, \vec{t}^{+}, \vec{t}^{-}\}$，并将其中 90%作为训练集，10%作为验证集。

在模型训练之前，我们首先使用文献[104]中训练好的神经网络参数对本模型进行初始化。在训练过程中，同文献[93]中介绍的一样，首先固定其他网络层的参数不变，单独训练 M 层。当模型逐渐稳定后，再将损失函数的值后验传播到其他所有网络层，对整个网络模型进行微调。尽管如此，因为视觉-语义嵌入模型需要依靠大规模数据来学习图像的视觉特征，进而挖掘视觉信息与语义信息之间的关联性，本节中使用的带有用户点击信息（click-through）的视频数据集的规模还远远不够。为了解决以上问题，我们将在下一节中介绍如何进行多任务的视觉-语义深度嵌入模型训练。

5.3 基于多任务的视觉-语义深度嵌入技术

多任务学习策略（Multi-task Learning Strategy）可以结合多个任务的训练样本，通过抽取和利用多任务之间的共享信息，达到同时提升这些任务学习效率的目的[105]。因此，多任务学习策略可以有效解决单一任务训练样本较少的问题，通过与其他相关任务结合，借助共同的训练样本融合，提高单一任务的学习效率。因此，在本节中，我们使用多任务学习策略，通过结合 Clickture 图像数据集来扩展原有数据集，以提升训练得到的深度-语义嵌入模型在测量文本查询与视频缩略图之间关系上的性能。Clickture 图像数据集[106]是一个大规模的带有用户点击信息的图像数据集，近期在图像检索领域得到

[13] “Natural Language Toolkit,” http://www.nltk.org/.

广泛引用[14]。尽管查询相关的视频缩略图生成和图像检索是两个不同的任务，他们之间还是存在很多联系，比如都是计算语义与视觉信息之间的相似性，并且都反映了用户查询意图和查询语句之间的关系。因此，我们可以使用 Clickture 图像数据集来进行基于多任务学习的视觉-语义深度嵌入学习。在本章的实现中，我们使用了 Clickture 图像数据集中的训练集，共包含 2,310 万个带有点击信息的三元组$\{query, image, click\}$，其中有 1,170 万个不同的查询语句和 100 万张不同的图像。经过和 5.2 节中类似的预处理，我们最终使用了 100 万个$\{v, t^+, t^-\}$ 三元组。

根据多任务学习的基本原理，我们可以将原本深度语义-嵌入模型的优化目标修改为公式(5-2)所示：

$$\min_{M_k} \tau_0 \left\| M_0 - I \right\|_F^2 + \sum_{k=1}^{2} \{ \tau_k \left\| \Delta M_k \right\|_F^2 + \max[0, \gamma - S(t_k^+, v) + S(t_k^-, v)] \} \tag{5-2}$$

公式中，M 仍然表示卷积神经网络中的映射层。不同的是，M_0 捕获到的是多个数据集的一般趋势，而 $M_k = M_0 + \Delta M_k$ 则更多倾向于每个单独的任务。因为多任务学习的一个重要准则就是对多个学习任务的有效整合，因此最小化 $\left\| M_0 - I \right\|_F^2$ 和 $\left\| \Delta M_k \right\|_F^2$ 是为了保证学习算法既不会只偏重于多任务之间的共享信息，又不会只专注于某一个单一任务。与公式(5-1)类似，最小化 $\max[0, \gamma - S(t_k^+, v) + S(t_k^-, v)], S(t_k, v) = \vec{t}_k M_k \vec{v}_k$ 是为了确保学习到的模型可以使得视频缩略图对应的语义特征向量与相关联查询的语义特征向量的点积相似度要大于与无关联文本的点积相似度。$\tau_k \geq 0, k = 0, 1, 2$ 是一个权衡参数，用来决定训练得到的模型是倾向于捕获多任务之间的共同趋势，还是更倾向于某个单独的任务。

基于多任务的视觉-语义深度嵌入模型的训练过程分为两部分：首先，我们使用随机梯度下降法在所有任务的数据样本上学习 M_0。与 5.2 节中不同的是，这里的查询语句t和视觉数据v是从图像和视频数据集中生成的，并且损失函数变成了如下形式：

$$loss(v, \vec{t}^+, \vec{t}^-) = \tau_0 \left\| M_0 - I \right\|_F^2 + \sum \max[0, \gamma - \vec{t}^+ M_0 \vec{v} + \vec{t}^- M_0 \vec{v}], \tag{5-3}$$

在这一步中，我们可以认为深度网络中的前七层抽取了图像和视频缩略图之间共同展现的视觉特征。同时，M_0 层通过挖掘文本查询和图像/缩略图之间的共现性，将视觉特征映射到了隐含的语义空间。

在训练得到 M_0 之后，我们需要在视频数据集上单独对 M_0 进行微调得到 M_1。在这一步中，我们同样固定网络的前七层，只在带有用户点击信息的视频数据集上训练 M_1 层。与第一步所不同的是，损失函数变成了

[14] “Clickture,” http://research.microsoft.com/en-us/projects/clickture/.

$$loss(v,\vec{t}^{+},\vec{t}^{-}) = \tau_1 \left\| M_1 - M_0 \right\|_F^2 + \sum \max[0, \gamma - \vec{t}^{+} M_1 \vec{v} + \vec{t}^{-} M_1 \vec{v}], \quad (5\text{-}4)$$

其中$\left\| M_1 - M_0 \right\|_F^2$是为了避免在视频数据集上发生过拟合，$\tau_1$同样是权衡参数。在训练开始时，我们可以给$\tau_1$赋予一个较大的值，如 1.0，然后在训练时根据测试误差不断对其进行微调。因为我们训练视觉-语义深度嵌入模型的目的是进行查询相关的视频缩略图选择，所以在本节中，我们只训练了针对该任务的参数矩阵M_1而忽略了针对图像检索任务的参数矩阵M_2。在以后的工作中，我们将会尝试在同一个卷积神经网络中同时针对多个任务的映射层M_k进行并行训练。

5.4 查询相关的视频缩略图选择算法

在本节中，我们将介绍如何利用训练得到的基于多任务的视觉-语义深度嵌入技术进行查询相关的视频缩略图选择。整个算法流程可以分为如下两个阶段：离线阶段和在线阶段。

尽管我们的目的是选择查询相关的视频缩略图，但是我们首先不能忽略视频缩略图的主要任务还是要展示视频的主要内容。因此，在离线阶段，我们会根据视频关键帧的视觉表现属性，挑选具有代表性的关键帧作为候选缩略图。受文献[78]启发，我们首先根据颜色直方图将视频分割为不同的场景、镜头和子镜头。然后对于一个子镜头中的关键帧，提取了视频关键帧的视觉表现属性。为了抽取视觉质量好的视频关键帧作为缩略图，我们抽取了以下视觉表现属性：1）关键帧所在子镜头、镜头和场景的时间长度属性，以及邻接子镜头的长度属性；2）视频关键帧的位置属性、颜色熵值属性、运动模糊属性和边缘锐度属性；3）邻接视频关键帧的相似度属性。为了选择更有视觉吸引力的关键帧作为视频缩略图，我们抽取了人脸属性和肤色比例属性作为视觉表现属性。之后，将这些属性值的线性组合作为视频关键帧的视觉代表性评分。最后，分数最高的 20 个视频关键帧被选择作为候选缩略图。候选缩略图选择完毕后，会通过训练得到的基于多任务学习的视觉-语义深度嵌入模型映射到隐含的语义空间，得到语义特征向量。

在线视频搜索阶段，当搜索引擎接收到用户的查询语句之后，我们同样也把查询文本映射到语义空间，得到语义特征向量。在隐含的语义空间，通过查询语句和视频缩略图的语义特征向量计算二者之间的相似性。因为一个查询语句可能包含多个单词，视频缩略图选择算法将会计算每个单词与候选缩略图之间的相似性，并选择最高的分数为每个候选缩略图的查询相关性评分。最后，我们将融合视觉代表性评分和查询相关性评分，将得分最高的视频缩略图作为移动视频检索结果的展示方法。

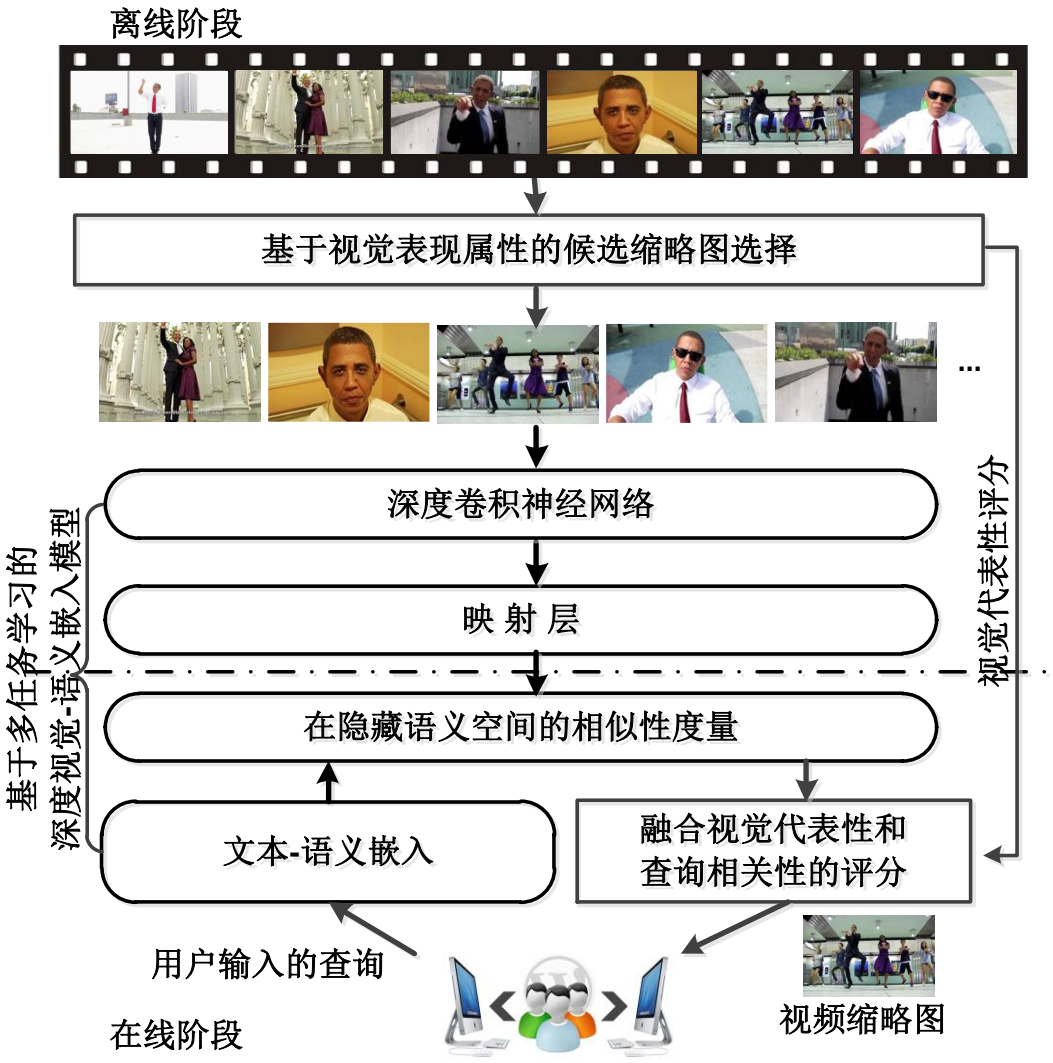


图 5-3 查询相关的视频缩略图选择算法流程图

5.5 实验结果

5.5.1 评测数据集

为了验证本文所提算法的实际性能，我们在带有用户点击信息的视频数据集中选择了 1000 个查询-视频组合。这些视频包含 9 个类别，平均长度为 331 秒，并且这些查询-视频组合并没有在训练集中出现。在这些查询-视频组合中，我们首先使用基于视觉表现属性的方法抽取了 17,480 个候选视频缩略图。因为视频缩略图的选择是一个主观评价任务，一个视频中通常有多帧图像都适合做缩略图，而且视频缩略图的好坏受到用户主观意识的影响。因此，为了使实验结果更加客观公正、降低标注成本，我们在亚马逊提供的 Amazon Mechanical Turk (AMT) 标注平台上，借助群体智慧实现视频缩略图质量

的标注[15]。为了保证标注质量不过于偏重于个别标注人员的主观意识，我们要求每个查询-视频组合都要被五个不同的标注人员标注，因此共发布了 5,000 个标注任务。每个标注任务包括一个查询-视频组合，20 张候选缩略图。标注人员必须首先了解查询语句，并观看视频，然后对候选缩略图进行评分。分数共设为如下五档：

1) 非常好(VG): 视频缩略图概括了视频的全部内容，图像的质量也非常好，并且和查询非常相关；
2) 好(G): 视频缩略图基本概括了视频的全部内容，并且与查询部分相关，但图像的质量不好；
3) 一般(F): 视频缩略图基本概括了视频的全部内容，或者与查询部分相关；
4) 坏(B): 视频缩略图的图像质量很好，但是既没有概括视频的全部内容，也和查询无关；
5) 非常坏(VB): 既没有概括视频的全部内容，图像质量也很差，并且和查询无关。

除此之外，为了控制标注质量，我们增加了如下需求：

1) 只有至少成功标注过 100 个任务并且标注认证率在 80%以上的标注人员才能参与此次标注；
2) 为了保证每个候选视频缩略图都被至少五个不同的标注人员标注，如果一次任务中有一个候选视频缩略图没有被标注，将不使用该次标注结果；
3) 对于 65%的视频，我们都选择了至少一个为 VB 或者 VG 的视频缩略图作为种子。如果一个标注人员有超过 3 次对种子缩略图的标注与实际值严重不符（比如对原本为 VB 种子标注了高于 B 的评分，或者对原本为 VG 的种子标注了低于 G 的评分），则该标注人员将禁止再次参与标注。
4) 如果标注人员对某个视频 90%以上的候选视频缩略图标注了同样的分数，我们将人工检查该视频，来决定是否使用该结果。

在加入以上要求后，我们共取消了 12%的标注任务结果，并将这些任务全部重新标注。经过 15 天的标注工作之后，共有超过 191 个标注人员参加了此次标注工程，每个标注任务平均耗时 355 秒。对于每个视频缩略图，我们选择了被标注最多的分数为最终分数，如果有两个分数被标注次数相同，则选择较低的分数。一些标注视频和结果的例子如图 5-6 所示。目前，该数据集已经公开发布[16]，作为查询相关的视频缩略图研究的公开评测数据集。

[15] “Amazon Mechanical Turk,” https://www.mturk.com.

[16] “查询相关视频缩略图选择评测数据集”，http://mcg.ict.ac.cn/mcg-vts.html

5.5.2 实验设置

为了对本文所提出的基于多任务学习的视觉-语义深度嵌入技术以及查询相关的视频缩略图选择算法进行评测，我们在评测数据集上测试了以下七种算法：

1) **RANDOM**：随机方法，该方法随机选择候选缩略图中的一帧图像作为最终的视频缩略图。
2) **ATTR**：基于视觉表现性的方法（Video Representation Attributes based Method）。该方法使用了文献[78]提出的视觉表现属性的方法，选择最具视觉代表性的关键帧作为视频缩略图。
3) **CCA**：基于典型相关性分析（Canonical Correlation Analysis，简称 CCA）的方法。典型相关性分析是一种经典的计算图像和文本相似度的方法[91][92]。该方法通过类似于主成分分析的方法，分别抽取图像和文本特征中具有代表性的变量，将其映射到隐含的空间，通过两组指标在隐藏空间中的相关关系，来代替原图像和文本之间的相关关系。我们通过文献[91]中提供的程序接口，在带有用户点击信息的图像和视频数据集上训练了 CCA 模型，并用来进行文本查询与视频缩略图之间的相似性度量。
4) **VSEM-VIDEO:** 基于视觉-语义深度嵌入技术的方法，嵌入模型只在带有点击信息的视频数据集上训练得到。
5) **VSEM-ALL:** 基于视觉-语义深度嵌入技术的方法，嵌入模型在带有点击信息的图像和视频数据集上训练得到。
6) **MTL-VSEM:** 基于多任务学习的视觉-语义深度嵌入技术的视频缩略图选择方法，嵌入模型通过多任务学习的策略在带有点击信息的图像和视频数据集上训练得到。
7) **FUSION:** 融合视觉代表性和查询相关性的方法。视觉代表性评分由 ATTR 得到，查询相关性评分由 MTL-VSEM 得到。

因为视频搜索引擎可以给用户提供一张或多张视频缩略图作为查询结果展示，因此我们使用两种指标作为评测标准：1）HIT@1，该指标只计算评分最高的视频缩略图的命中率；2）平均准确率（Mean Average Precision, MAP），该指标计算了所有视频缩略图的选择准确率。MAP 的计算公式如(5-5)所示：

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \tag{5-5}$$

公式中，Q是查询集合，m_j是每个查询视频对中有效视频缩略图的数量，$Precision(R_{jk})$是在算法返回第k个有效缩略图时的准确率。因为我们使用了五个不同的分数来标注视频缩略图，因此我们分两种情况来定义有效视频缩略图：1）标注为 VG 的缩略图；2）

标注为 VG 或 G 的视频缩略图。

5.5.3 在完整数据集上的性能评测

我们首先在标注好的全部 1000 个查询-视频组合上比较了以上 7 种方法的性能。HIT@1 的结果如表格 5-1 所示。而 MAP 指标评测的结果如图 5-4 所示。从结果中可以

表格 5-1 不同视频缩略图选择算法的性能评测（使用 HIT@1 指标）。

Method	HIT@1(VG)	HIT@1(VG&G)
RANDOM	26.95%	55.68%
ATTR [78]	40.21%	68.89%
CCA [91]	31.66%	60.06%
VSEM-VIDEO	32.96%	62.48%
VSEM-ALL	39.88%	67.84%
MTL-VSEM	**43.03%**	**71.70%**
FUSION	**47.13%**	**74.83%**

看到，无论是只选择一个缩略图，还是选择多个缩略图，我们的方法都在七个方法中获得了最高的准确率。与 ATTR 方法的结果相比，我们的方法可以明显提高视频缩略图选择的质量。该结果证明了我们的查询相关的视频缩略图选择方法确实能够捕获用户的查询意图，返回更能代表视频内容的视频缩略图。与 CCA 方法的结果相比，我们的方法同样性能更好。这是因为 CCA 的方法只是训练了一个线性转换矩阵，用来将视觉和文本特征映射到隐含的空间，来计算二者之间的相似性。但是，视觉和文本特征并不一定是线性相关的。与 CCA 方法不同的是，在视觉-语义嵌入模型中，所有的误差都会通过后向传播到达网络的所有层次，从而可以直接对视觉特征进行优化，尽可能多的挖掘图像和文本之间的相似关系。

更重要的是，在所有基于视觉语义嵌入技术的查询相关视频缩略图生成算法中（VSEM-VIDEO, VSEM-ALL 和 MTL-VSEM），本文提出的 MTL-VSEM 方法仍然取得了最高的准确率，其原因如下：VSEM-VIDEO 方法只在有限的视频数据集上进行训练，学习到的视觉-语义嵌入模型并不能有效挖掘视觉-语义之间的关系。尽管增加了图像数据集，VSEM-ALL 方法并没有充分考虑图像检索和视频缩略图选择这两个不同的任务之间的不同之处。而且由于图像数据集规模要远远大于视频数据集，所训练得到的视觉-语义嵌入模型往往将重点过多的放在了图像检索上，而没有针对视频缩略图选择任务进行微调优化。与前两种方法不同，MTL-VSEM 方法既有效避免了模型过多依赖于图

像数据集，又避免了在视频数据集中的过拟合。最后，融合视觉代表性和查询相关性的方法取得了最好的性能，证明了视觉表现性和查询相关性都是视频缩略图选择的关键。

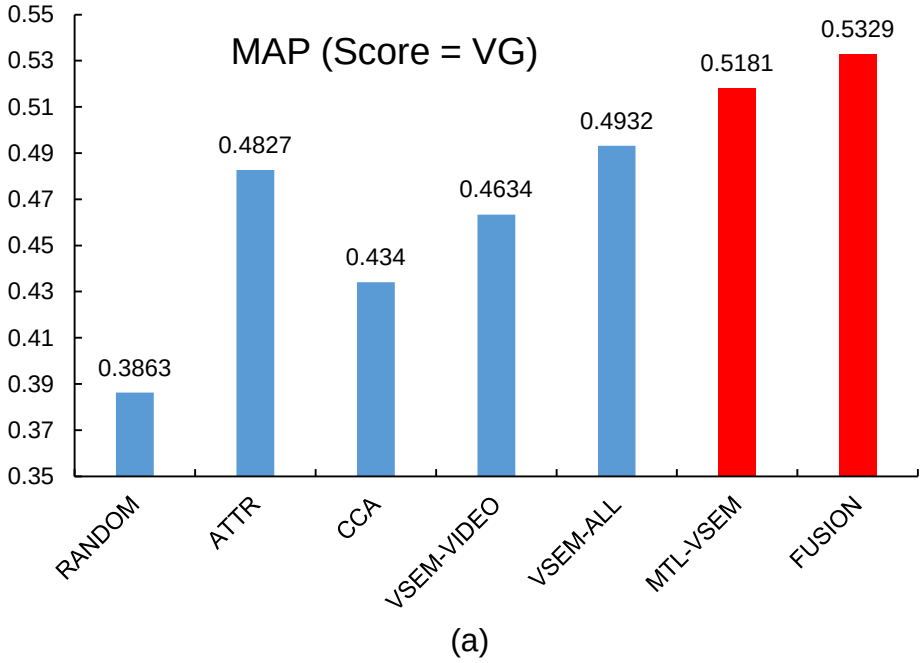


图 5-4 不同视频缩略图选择算法的评测结果（使用 MAP 指标）

5.5.4 在不同视频类别上的性能评测

为了进一步对本文所提方法进行评测，我们在 9 个不同视频类别上对相关方法进行了评测。9 个视频类别分别是：教育，休闲，电影，游戏，音乐，新闻，物体，人物和运动。我们使用 MAP 作为评价指标，缩略图的有效分数为 VG 和 G，参与评测的方法

有 ATTR，VSEM-ALL，MTL-VSEM 和 FUSION 四种方法。

最终的实验结果如图 5-5 所示。从图中可以发现，MTL-VSEM 和 FUSION 方法的性能在所有类别上都要优于其他方法。特别是在“物体”和“教育”类视频上，性能提升非常明显。该结果证明了当用户搜索这两类视频时，其目的性要更强。与此相对应的是，在搜索“运动”和“电影”类视频时，基于视觉表现性的方法和基于查询相关性的方法都取得了相似的结果，其原因是在这类视频中，大部分视频关键帧内容都与主题非常相关。比如在足球比赛的视频中，全部视频内容都是跟足球比赛非常相关的，因此基

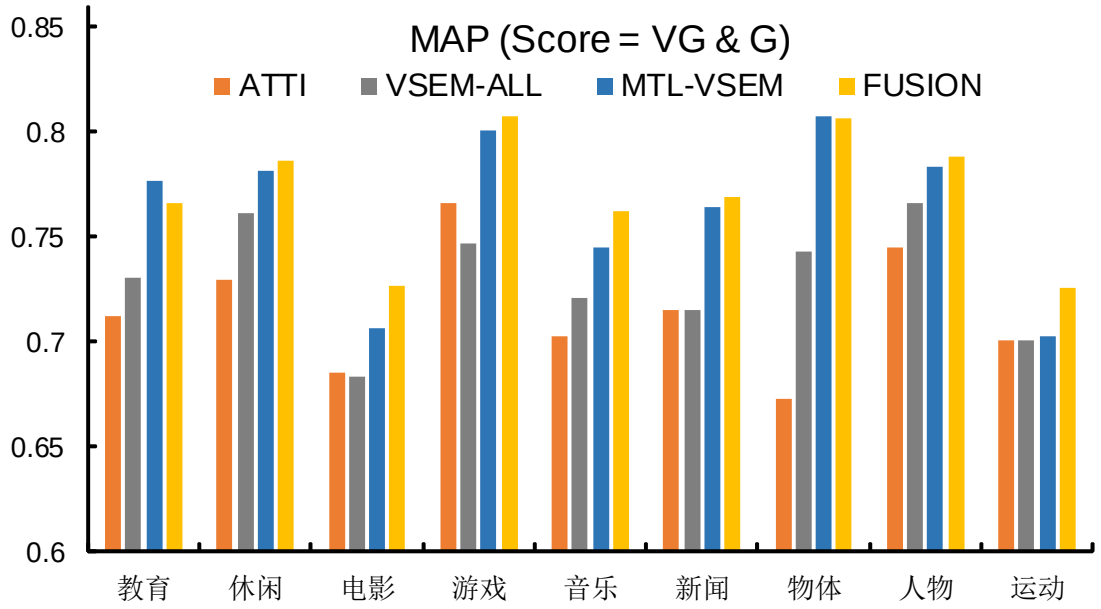


图 5-5 不同视频缩略图选择算法在 9 种视频类别上的评测结果（使用 MAP 指标）

于视觉表现性的方法同样能容易找到和查询相关的视频关键帧。但是，由于两种视频缩略图算法都很难捕获到视频中的精彩镜头或者某个特定的演员，因此这两类方法在“运动”和“电影”类视频中都不能取得特别令人满意的结果。除此之外，与 VSEM-ALL 方法相比，MTL-VSEM 方法同样在所有类别上都取得了更好的结果，特别是针对“物体”和“游戏”类视频，该结果进一步证明了多任务学习策略的有效性。

为了让读者对不同视频缩略图选择方法的实际效果有一个更加直观的认识，我们在图 5-6 中展示了不同方法在不同视频类别上视频缩略图选择的结果。从该结果中可以非常直观的看到，与其他比较方法相比，VSEM-ALL 和 FUSION 方法都选择了既能代表视频内容，又与用户输入查询高度相关的视频缩略图。这些视频缩略图能够帮助用户快速定位目标视频，提高搜索体验。作为补充，我们还在图 5-6（b）中给出了两个选择效果不是很好的例子。在第一个例子中，VSEM-ALL 和 FUSION 方法都选择了橄榄球比赛的画面，而不是查询中提到的橄榄球运动员“Lance Briggs”的特写。在第二个例

子中，两个方法同样没能找到演员 Matt Damon 的特写。其原因也是因为由于两种视频缩略图算法都很难捕获到视频中的精彩镜头或者某个特定的演员。

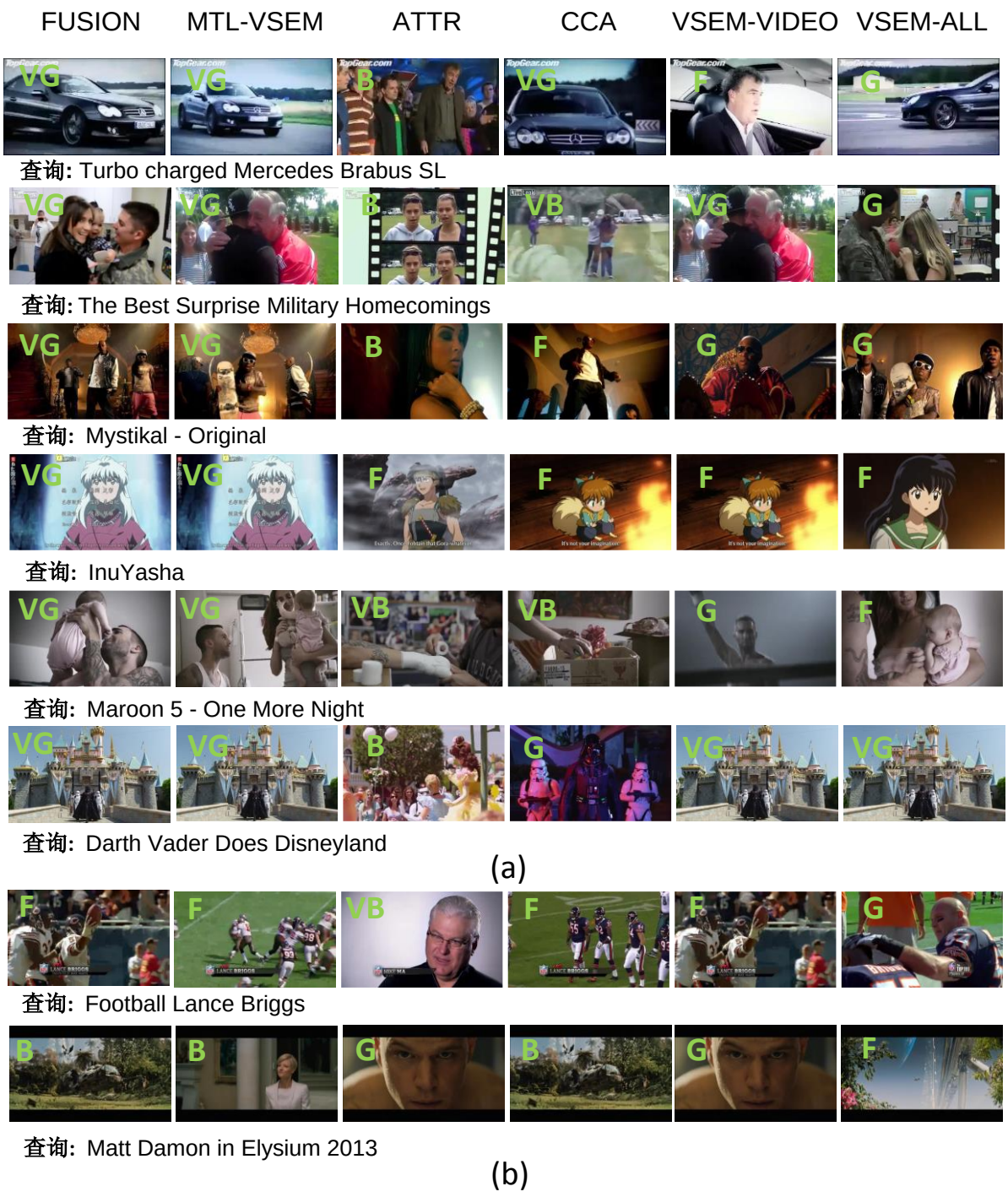


图 5-6 不同视频缩略图选择算法结果示例图。（a）成功的例子;（b）失败的例子

5.6 小结

在本章中，针对移动视频检索的结果展示问题，我们研究设计了查询相关的视频缩

略图选择算法，根据用户输入的文本查询，自动生成满足用户查询意图的视频缩略图，帮助用户快速在移动设备上发现目标视频。我们通过研究视觉-语义深度嵌入技术，将文本查询与高视觉代表性的视频关键帧同时映射到隐含的语义空间，并在该空间中计算文本与视觉信息的相似度。该方法不受训练集语义类别的限制，能够有效计算多类别查询与视频关键帧的相似性。同时，针对视频训练数据集规模不足的问题，我们研究融合了规模更大的图像检索数据集。通过结合多任务学习的策略，将在共同数据集中学习到的知识迁移到视频缩略图选择中来，消除图像与视频之间的障碍，训练得到了性能更好的视觉-语义嵌入模型。最后，通过融合视频关键帧的视觉代表性和查询相关性评分，选择了最能代表视频视觉和语义信息的视频缩略图。实验证明，本文研究的查询相关的视频缩略图选择算法，能够帮助移动用户快速了解视频内容，显著提高用户的查询体验。

尽管在本章中，基于多任务学习的视觉-语义嵌入模型只用来进行查询相关的视频缩略图选择，但是该模型同样可以用于其他任务，比如视频标签定位[107]、视频摘要生成[108]、视频自动语言描述[109]等。在以后的工作中，我们将对这些应用进行进一步的研究。除此之外，我们还将研究如果通过加入更多的映射层 M_k 来同时训练多个不同的任务，并且尝试结合视频的字幕、自动语音识别和人脸识别等来选择更满足用户查询意图的视频缩略图。

第6章 结束语

6.1 工作总结

移动互联网、智能移动设备和多媒体技术的高速发展改变了互联网视频内容的产生以及人们检索和观看视频的方式。越来越多的人希望在移动中接入互联网进行视频搜索，便捷地获取全面的视频信息及服务。与传统视频检索不同，移动设备无处不在的网络接入能力使得移动用户随时随地都能通过移动设备来检索、观看和分享视频数据；移动设备上丰富的传感器资源能够快捷贴切地反映和交互修正用户的查询意图，有利于克服用户查询意图和查询表示的语义鸿沟，并通过丰富的上下文信息辅助视频检索，提升性能和效率；同时，移动设备不断增长的计算能力能够快速的处理用户输入的查询，实时返回搜索结果，提升搜索体验。但另一方面，移动视频检索仍然面临查询视频受外界干扰严重，网络视频数据规模庞大、质量参差不齐，移动设备计算性能与桌面 PC 仍然存在差距，移动网络带宽限制和移动用户对移动搜索体验的特殊要求等诸多问题。研究怎样解决这些问题，提供快速、准确的移动视频检索技术，满足用户特有的移动视频搜索需求，成为移动视频检索的主要研究内容。因此，移动视频检索的相关研究，不仅具有广阔的应用前景和巨大的市场价值，也为移动社交网络、移动互联网电子商务、地标识别、实景导航和增强现实交互等移动互联网应用提供了关键技术支持，推动移动互联网新一代技术的突破。同时，移动视频检索关键技术研究涉及计算机视觉、图像处理、机器学习、模式识别和信息检索等相关领域重要基本问题的攻克，是一个极富挑战性的研究方向。

综上所述，本篇论文着重研究了移动视频检索的若干关键技术，包括移动视频签名生成与加速算法、音/视频分层哈希索引与匹配算法和查询相关的视频缩略图生成算法。我们首先对移动视频检索相关技术进行了较为全面的综述和概括，包括传统近似视频检索和现有移动视觉检索中的相关概念、发展历史和相关方法，并分析总结了传统视频检索技术应用于移动视频检索时的不足。接下来，我们从移动视频检索对视频签名生成、视频签名索引和匹配以及视频检索结果展示的新要求出发，围绕移动视频签名的提取、压缩、加速，移动视频的快速索引和精确匹配，以及查询相关的视频缩略图生成技术进行了较为深入的研究，显著降低了移动视频检索的存储、传输和计算开销，提高了移动视频检索的精度、速度和用户搜索体验。本文的主要研究成果总结如下：

1. 移动视频签名生成与加速算法

视频签名是对一个视频对象的感知特征或简短的摘要，现有签名技术通常具有无冲突性、安全性、紧凑性、鲁棒性和篡改敏感性等特征。除此之外，由于移动视频检索的

特有挑战，还要求移动视频签名技术具有计算简单、区分性强、易传输和易索引等特性。基于以上原因，我们分别提取了基于渐进式传输的视觉哈希码和基于频域局部显著性的音频指纹特征作为移动视频签名。借助音/视频信号的互补性，本文提出的视频签名可以有效应对移动视频检索中查询视频因拍摄环境造成的音/视频信号改变。通过使用谱哈希技术进行二进制编码，显著降低了需要传输的特征规模，提高了传输和索引的速度。除此之外，本文提出的基于 Hessian 显著度加权融合的渐近式传输方法，可以根据不同的移动设备性能和网络环境，自适应的选择需要传输的特征规模，进一步提高移动视频检索的识别比特率，保证了检索的实时性。为了验证本文所提视频签名的性能，我们搜集并发布了世界上第一个基于真实应用环境的移动视频检索数据集 MCG-MVQ。在该数据集上的实验表明，在主流移动设备上，本文所提出的移动视频签名的提取时间仅为 267ms，每秒传输数据量只有 0.88 KB。与已有方法相比，获得了最高的识别比特率。

2. 音/视频分层哈希索引与匹配算法

目前已有的二进制哈希索引算法都是针对单一类型的特征进行处理，其性能非常有限。针对移动视频检索丰富的模态查询信息，本文研究从多种特征模态入手挖掘特征之间的关联，实现了音/视频分层哈希索引与匹配方法。通过音/视频分层过滤策略显著提高了视频签名检索的速度；通过两次不同的融合过程，充分挖掘了音/视频特征间的互补性，显著提高了视频签名的识别能力和检索速度。除此之外，针对移动视频检索子序列匹配任务的特点，我们提出了利用二分图转换和最大匹配技术实现查询视频与源视频的渐进式匹配，在提高查询准确率的基础上，显著提高了用户的使用体验。最后，我们结合本文所提出的视频签名生成与加速方法、视频签名索引与匹配方法，设计并实现了一个完整的移动视频检索系统。经实验验证，在 MCG-MVQ 数据集上，该系统在查询视频短于 10 秒的基础上获得 91.59%以上的准确率。经易用性主观评测证明，与已有移动视频检索系统相比，该系统不仅检索性能更高，还具有更好的用户使用体验。

3. 基于多任务学习的深度-视觉语义嵌入技术

视频缩略图能集中展现视频的主要内容，帮助人们快速了解检索结果。在移动视频检索中，网络带宽和流量限制了用户快速浏览视频内容，视频缩略图就成为用户了解视频内容，确定目标视频的最有效方式。但是现有视频缩略图选择方法没有考虑视频语义信息，无法体现用户的查询意图。本文中，我们研究了通过视频附加语义信息选择查询相关的视频缩略图方法。该方法首选使用基于多任务学习策略的视觉-语义深度嵌入技术将文本信息和视觉信息映射到隐含的语义空间，并在该空间中计算他们之间的距离。通过多任务学习的策略，联合学习多任务之间的中间态参数模型，可以在大规模带有用户点击信息的视频和图像数据集上训练有效的嵌入模型。最后，借助群体智慧，我们在亚马逊 AMT 标注平台上邀请了 191 名标注人员参与标注了 17,480 个查询-视频缩略图组合，并在此数据集上验证了本文方法的性能。实验证明，与现有最好的方法相比，用户对本文方法生成的视频缩略图的满意度提高了 6%。该结果证明本文所设计的查询相

关视频缩略图选择算法不仅能有效概括视频核心内容，还可以通过视觉-语义深度嵌入技术捕获用户查询意图，生成个性化的移动视频检索结果，帮助移动用户快速发现目标视频。

6.2 研究展望

本文对移动视频检索中的若干关键问题进行了较为深入的研究，并取得了一定的研究成果，但是，总的来说，移动视频检索还是一个新兴的研究方向，仍然存在很多问题需要解决。值得进一步研究的问题包括：

1. 多模态的视频检索方式

目前常见的移动视频检索主要是通过文本和通过视频录制两种方式。每种方式都有其局限性：比如由于存在语义鸿沟，文字输入无法准确表现用户查询意图；视频录制方式必须录制正在播放的视频等。因此，如何结合文本输入和视频录制两种方式，设计更为丰富、合理的交互模态、输入方式，使得用户交互更加丰富，帮助用户更好的表达检索意图，值得进一步研究。

2. 分布式移动视频索引与匹配

面对互联网海量视频数据，单台服务器已无法满足高效存储及检索的需求。因此，如何结合本文所提出的移动视频索引和匹配技术，针对分布式计算的特点，设计基于分布式计算的视频索引与匹配算法，建立高效的索引机制，在大规模机群上进行视频索引的部署和查询，是一个非常值得研究的学术方向，并且具有重要的应用价值。

3. 基于多任务的视觉-语义深度嵌入

本文在视觉-语义深度嵌入模型训练中使用的多任务学习策略，在挖掘多任务间共同模式的基础上，只针对单一任务进行了优化处理。怎样利用深度神经网络模型的特点，在同一网络模型中同时对多个子任务并行优化，可以进一步提高优化的效率和性能。除此之外，研究如何将视觉-语义深度嵌入模型应用到视频摘要生成、视频标签定位和视频自动语言描述等问题中，也具有重要的学术和应用价值。

4. 移动设备本地化应用

由于移动设备往往与个人信息紧密绑定，我们可以结合用户使用习惯和社交网络，为用户提供个性化的搜索结果。除此之外，研究如何将移动视频检索与视频推荐、视频分享、智能广告推荐、智能电视等研究问题结合，实现跨模态、跨媒体检索与推荐，将赋予移动视频检索新的生命。

参考文献

[1] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, “CHoG: Compressed histogram of gradients A low bit-rate feature descriptor,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2504 –2511.

[2] 工信部电信研究院通信信息所数据监测部, “我国移动互联网发展情况分析,” 人民邮电报.

[3] 段凌宇, 黄铁军, and 高文, “移动视觉搜索技术研究与标准化进展,” 信息通信技术, 2012.

[4] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, “Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing,” in *ACM Multimedia*, 2012, pp. 9–18.

[5] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, “Mobile product search with Bag of Hash Bits and boundary reranking,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3005–3012.

[6] J. Huber, J. Steimle, and M. Mühlhäuser, “Mobile interaction techniques for interrelated videos,” in *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, 2010, pp. 3535–3540.

[7] 唐胜, “多媒体数字签名技术研究,” 博士学位论文, 中国科学院计算技术研究所, 2006.

[8] Y.-C. Su, T.-H. Chiu, Y.-Y. Chen, C.-Y. Yeh, and W. H. Hsu, “Enabling Low Bitrate Mobile Visual Recognition: A Performance Versus Bandwidth Evaluation,” in *ACM Multimedia*, 2013, pp. 73–82.

[9] X. Yang and K.-T. T. Cheng, “Accelerating SURF Detector on Mobile Devices,” *ACM Multimedia*, 2012, pp. 569-578.

[10] J. Xia, K. Gao, D. Zhang, and Z. Mao, “Geometric context-preserving progressive transmission in mobile visual search,” in *ACM Multimedia*, 2012, pp. 953–956.

[11] V. R. Chandrasekhar, S. S. Tsai, G. Takacs, D. M. Chen, N.-M. Cheung, Y. Reznik, R. Vedantham, R. Grzeszczuk, and B. Girod, “Low latency image retrieval with progressive transmission of CHoG descriptors,” in *Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing*, 2010, pp. 41–46.

[12] R. Ji, L.-Y. Duan, J. Chen, H. Yao, Y. Rui, S.-F. Chang, and W. Gao, “Towards low bit rate mobile visual search with multiple-channel coding,” in *ACM Multimedia*, 2011, pp.

573–582.

[13] K.-Y. Tseng, Y.-L. Lin, Y.-H. Chen, and W. H. Hsu, “Sketch-based image retrieval on mobile devices using compact hash bits,” in *ACM Multimedia*, 2012, pp. 913–916.

[14] R. Yan, J. Yang, and A. G. Hauptmann, “Learning query-class dependent weights in automatic video retrieval,” in *ACM Multimedia*, 2004, pp. 548–555.

[15] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, “Video Search Reranking via Information Bottleneck Principle,” in *ACM Multimedia*, 2006, pp. 35–44.

[16] A. Wang, “An industrial strength audio search algorithm,” in *Proc. Int. Conf. on Music Info. Retrieval ISMIR*, 2003, vol. 3.

[17] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, “Coherent bag-of audio words model for efficient large-scale video copy detection,” in *ACM Multimedia*, 2010, pp. 89–96.

[18] H. Jegou, J. Delhumeau, J. Yuan, G. Gravier, and P. Gros, “BABAZ: A large scale audio search system for video copy detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2369 –2372.

[19] J. Sivic and A. Zisserman, “Video Google: A Text Retrieval Approach to Object Matching in Videos,” in *IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[20] L. Shang, L. Yang, F. Wang, K.-P. Chan, and X.-S. Hua, “Real-time large scale near-duplicate web video retrieval,” in *ACM Multimedia*, 2010, pp. 531–540.

[21] M. Douze, H. Jégou, C. Schmid, and P. Pérez, “Compact video description for copy detection with precise temporal alignment,” in *PEuropean conference on Computer vision: Part I*, 2010, pp. 522–535.

[22] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, “Early versus late fusion in semantic video analysis,” in *ACM Multimedia*, 2005, pp. 399–402.

[23] Y. Liu, C. Xu, and H. Lu, “Audio-visual large-scale video copy detection,” *International Journal of Computer Mathematics*, vol. 88, no. 18, pp. 3803–3816, 2011.

[24] W. Liu, T. Mei, Y. Zhang, J. Li, and S. Li, “Listen, look, and gotcha: instant video search with mobile phones by layered audio-video indexing,” in *ACM Multimedia*, 2013, pp. 887–896.

[25] W. Liu, T. Mei, and Y. Zhang, “Instant Mobile Video Search With Layered Audio-Video Indexing and Progressive Transmission,” *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2242–2255, 2014.

[26] W. Liu, F. Yang, Y. Zhang, Q. Huang, and T. Mei, “LAVES: an instant mobile video search system based on layered audio-video indexing,” in *ACM Multimedia*, 2013, pp.

409–410.

[27]H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-Up Robust Features (SURF),” *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, 2008.

[28] Z. Huang, B. Hu, H. Cheng, H. T. Shen, H. Liu, and X. Zhou, “Mining near-duplicate graph for cluster-based reranking of web video search results,” *ACM Trans. Inf. Syst.*, vol. 28, no. 4, p. 22, 2010.

[29] X. Wu, A. G. Hauptmann, and C.-W. Ngo, “Practical elimination of near-duplicates from web video search,” in *ACM Multimedia*, 2007, pp. 218–227.

[30] L. Liu, W. Lai, X.-S. Hua, and S.-Q. Yang, “Video Histogram: A Novel Video Signature for Efficient Web Video Duplicate Detection,” in *3th International Multimedia Modeling Conference*, 2007, vol. 4352, pp. 94–103.

[31] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan, “Real-Time Near-Duplicate Elimination for Web Video Search With Content and Context,” *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 196–207, 2009.

[32] W. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, “Near-Duplicate Keyframe Identification With Interest Point Matching and Pattern Learning,” *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 1037–1048, 2007.

[33] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan, “Near-duplicate keyframe retrieval by semi-supervised learning and nonrigid image matching,” *TOMCCAP*, vol. 7, no. 1, p. 4, 2011.

[34] W. Liu, T. Xia, J. Wan, Y. Zhang, and J. Li, “RGB-D Based Multi-attribute People Search in Intelligent Visual Surveillance,” in *Advances in Multimedia Modeling*, vol. 7131, 2012, pp. 750–760.

[35] Y.-G. Jiang and C.-W. Ngo, “Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval,” *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 405–414, 2009.

[36] S. Poullot, M. Crucianu, and O. Buisson, “Scalable mining of large video databases using copy detection,” in *ACM Multimedia*, 2008, pp. 61–70.

[37] M. Douze, H. Jegou, and C. Schmid, “An Image-Based Approach to Video Copy Detection With Spatio-Temporal Post-Filtering,” *IEEE Transactions on Multimedia*, vol. 12, no. 4, pp. 257–266, Jun. 2010.

[38] Z. Huang, H. T. Shen, J. Shao, B. Cui, and X. Zhou, “Practical Online Near-Duplicate Subsequence Detection for Continuous Video Streams,” *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 386–398, 2010.

[39] Y. Yan, B. C. Ooi, and A. Zhou, “Continuous Content-Based Copy Detection over

Streaming Videos," in *Proceedings of the 24th International Conference on Data Engineering*, 2008, pp. 853–862.

[40] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A Review of Audio Fingerprinting," *J VLSI Sign Process Syst Sign Image Video Technol*, vol. 41, no. 3, pp. 271–284, Nov. 2005.

[41] J. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprinting System," in *3rd International Conference on Music Information Retrieval, ,* 2002.

[42] S. Baluja and C. Michele, "Content fingerprinting using wavelets," 2006, pp. 198–207.

[43] A. Wang, "The Shazam music recognition service," *Commun. ACM*, vol. 49, no. 8, pp. 44–48, 2006.

[44] R. Aly, "Modeling Representation Uncertainty in Concept-based Multimedia Retrieval," *SIGIR Forum*, vol. 44, no. 2, pp. 82–82, 2011.

[45] M. Hadjieleftheriou, Y. Manolopoulos, Y. Theodoridis, and V. J. Tsotras, "R-Trees - A Dynamic Index Structure for Spatial Searching," in *Encyclopedia of GIS.*, 2008, pp. 993–1002.

[46] T. K. Sellis, N. Roussopoulos, and C. Faloutsos, "The R+-Tree: A Dynamic Index for Multi-Dimensional Objects," in *International Conference on Very Large Data Bases*, 1987, pp. 507–518.

[47] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles," in A*CM International Conference on Management of Data*, 1990, pp. 322–331.

[48] V. Ramasubramanian and K. K. Paliwal, "Fast K-dimensional tree algorithms for nearest neighbor search with application to vector quantization encoding," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 518–531, 1992.

[49] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[50] P. Indyk and R. Motwani, "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing*, 1998, pp. 604–613.

[51] A. Gionis, P. Indyk, and R. Motwani, "Similarity Search in High Dimensions via Hashing," in *International Conference on Very Large Data Bases*, 1999, pp. 518–529.

[52] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, no. 1, pp. 117–122, 2008.

[53] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Multi-probe LSH: efficient

indexing for high-dimensional similarity search," in *Proceedings of the 33rd international conference on Very large data bases*, 2007, pp. 950–961.

[54] H. Jégou, M. Douze, and C. Schmid, "Product Quantization for Nearest Neighbor Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117 –128, Jan. 2011.

[55] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in NIPS, PP.1753-1760, 2008.

[56] R. Salakhutdinov and G. E. Hinton, "Semantic hashing," *Int. J. Approx. Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[57] G. Shakhnarovich, P. A. Viola, and T. Darrell, "Fast Pose Estimation with Parameter-Sensitive Hashing," in *IEEE International Conference on Computer Vision*, 2003, pp. 750–759.

[58] B. Kulis and T. Darrell, "Learning to Hash with Binary Reconstructive Embeddings," in *Advances in Conference on Neural Information Processing Systems 2009.*, 2009, pp. 1042–1050.

[59] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, "Spherical hashing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2957–2964.

[60] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 1–8.

[61] K. He, F. Wen, and J. Sun, "K-means Hashing: an Affinity-Preserving Quantization Method for Learning Binary Compact Codes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2938–2945.

[62] X. Zhang, L. Zhang, and H.-Y. Shum, "QsRank: Query-sensitive hash code ranking for efficient \unicode8714-neighbor search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2058–2065.

[63] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "AnnoSearch: Image Auto-Annotation by Search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1483–1490.

[64] J. Wan, S. Tang, Y. Zhang, L. Huang, and J. Li, "Data driven multi-index hashing," in *IEEE International Conference on Image Processing*, 2013, pp. 2670–2673.

[65] M. Muja and D. G. Lowe, "Fast Matching of Binary Features," in *2012 Ninth Conference on Computer and Robot Vision (CRV)*, 2012, pp. 404 –410.

[66] J. Yuan, L.-Y. Duan, Q. Tian, S. Ranganath, and C. Xu, "Fast and robust short video clip search for copy detection," *Advances in Multimedia Information Processing-PCM 2004*, pp. 479–488, 2005.

[67] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy

detection," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 15, no. 1, pp. 127–132, 2005.

[68] A. Hampapur, K. Hyun, and R. M. Bolle, "Comparison of sequence matching techniques for video copy detection," in *Storage and Retrieval for Media Databases* , 2002, vol. 4676, pp. 194–201.

[69] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video copy detection: a comparative study," in *ACM international conference on Image and video retrieval*, 2007, pp. 371–378.

[70] W.-L. Zhao, X. Wu, and C.-W. Ngo, "On the Annotation of Web Videos by Efficient Near-Duplicate Search," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 448 –461, Aug. 2010.

[71] Z. Huang, L. Wang, H. T. Shen, J. Shao, and X. Zhou, "Online Near-Duplicate Video Clip Detection and Retrieval: An Accurate and Fast System," in *IEEE International Conference on Data Engineering*, 2009, pp. 1511 –1514.

[72] S.-C. S. Cheung and A. Zakhor, "Efficient video similarity measurement with video signature," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 13, no. 1, pp. 59–74, 2003.

[73] 刘红, "近重复视频检测算法研究," 博士学位论文, 复旦大学, 2012.

[74] T. Can and P. Duygulu, "Searching for repeated video sequences," in *ACM International Workshop on Multimedia Information Retrieval* 2007, pp. 207–216.

[75] E. Maani, S. A. Tsaftaris, and A. K. Katsaggelos, "Local feature extraction for video copy detection in a database," in *Proceedings of the International Conference on Image Processing*, 2008, pp. 1716–1719.

[76] H. T. Shen, J. Shao, Z. Huang, and X. Zhou, "Effective and Efficient Query Processing for Video Subsequence Identification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 3, pp. 321 –334, Mar. 2009.

[77] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua, "Scalable detection of partial near-duplicate videos by visual-temporal consistency," in *ACM Multimedia*, 2009, pp. 145–154.

[78] H.-W. Kang and X.-S. Hua, "To learn representativeness of video frames," in *ACM Multimedia*, 2005, pp. 423–426.

[79] J. Luo, C. Papin, and K. Costello, "Towards Extracting Semantically Meaningful Key Frames From Personal Video Clips: From Humans to Computers," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 19, no. 2, pp. 289–301, 2009.

[80] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1346–1353.

[81] A. Rav-Acha, Y. Pritch, and S. Peleg, “Making a Long Video Short: Dynamic Video Synopsis,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 435–441.

[82] Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li, “Accurate Estimation of Human Body Orientation From RGB-D Sensors,” *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1442–1452, Oct. 2013.

[83] Z. Lu and K. Grauman, “Story-Driven Summarization for Egocentric Video,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2714–2721.

[84] A. Hanjalic, C. Kofler, and M. Larson, “Intent and its discontents: the user at the wheel of the online video search engine,” in *ACM Multimedia*, 2012, pp. 1239–1248.

[85] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, “Enriching and localizing semantic tags in internet videos,” in *ACM Multimedia*, 2011, pp. 1541–1544.

[86] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, “Large-Scale Video Summarization Using Web-Image Priors,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2698–2705.

[87] W. Zhang, C. Liu, Z. Wang, G. Li, Q. Huang, and W. Gao, “Web video thumbnail recommendation with content-aware analysis and query-sensitive matching,” *Multimed Tools Appl*, pp. 1–25, Jul. 2013.

[88] H. Li, L. Yi, B. Liu, and Y. Wang, “Localizing relevant frames in web videos using topic model and relevance filtering,” *Mach. Vis. Appl.*, vol. 25, no. 7, pp. 1661–1670, 2014.

[89] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, “Event driven web video summarization by tag localization and key-shot identification,” *Multimedia, IEEE Transactions on*, vol. 14, no. 4, pp. 975–985, 2012.

[90] S. Sun, “A survey of multi-view machine learning,” *Neural Comput & Applic*, vol. 23, no. 7–8, pp. 2031–2038, Dec. 2013.

[91] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, “A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics,” *Int J Comput Vis*, vol. 106, no. 2, pp. 210–233, Jan. 2014.

[92] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui, “Click-through-based cross-view learning for image search,” in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014, pp. 717–726.

[93] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, and others, “Devise: A deep visual-semantic embedding model,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2121–2129.

[94] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, “BRIEF: Computing a local binary descriptor very fast,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1281–1298, 2012.

[95] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: an efficient alternative to SIFT or SURF,” in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.

[96] Y. Wu, S. Lu, T. Mei, J. Zhang, and S. Li, “Local visual words coding for low bit rate mobile visual search,” in *ACM Multimedia*, 2012, pp. 989–992.

[97] Y. Gong and S. Lazebnik, “Iterative quantization: A procrustean approach to learning binary codes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 817–824.

[98] “Fourier transform,” *Wikipedia, the free encyclopedia*. 29-Mar-2015.

[99] M. Norouzi and D. Fleet, “Minimal Loss Hashing for Compact Binary Codes,” in International Conference on Machine Learning, 2011, pp.353-360.

[100] S. S. Tsai, D. M. Chen, G. Takacs, V. Chandrasekhar, J. P. Singh, and B. Girod, “Location coding for mobile image retrieval,” in *International Conference on Mobile Multimedia Communications, 2009*.

[101] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global Vectors for Word Representation,” in *Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pp. 1532–1543.

[102] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[103] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O’Reilly, 2009.

[104] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, “Deep Learning for Content-Based Image Retrieval: A Comprehensive Study,” in *ACM Multimedia*, 2014, pp. 157–166.

[105] J. Zhou, J. Chen, and J. Ye, “MALSAR: Multi-task learning via structural regularization,” *Arizona State University*, 2011.

[106] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li, “Clickage: towards bridging semantic and intent gaps via mining click logs of search engines,” in *ACM Multimedia*, 2013, pp. 243–252.

[107] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, “Discriminative segment annotation in weakly labeled video,” in *IEEE Conference on Computer Vision and Pattern*

Recognition, 2013, pp. 2483–2490.

[108] B. Zhao and E. P. Xing, “Quasi Real-Time Summarization for Consumer Videos,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2513–2520.

[109] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko, “Translating Videos to Natural Language Using Deep Recurrent Neural Networks,” *CoRR*, vol. abs/1412.4729, 2014.

致 谢

值此论文完成之际，首先向我的导师李锦涛研究员致以最诚挚的谢意。感谢李老师在过去的六年里对我的悉心栽培和孜孜教诲。“教诲如春风，师恩似海深”，从科研习惯培养，到博士选题与本论文的撰写与完成，李老师都倾注了大量的心血。除此之外，李老师更教会了我很多待人处事的道理。李老师渊博的知识、敏锐的洞察力、严谨的治学态度和高尚的学术道德都使我受益匪浅，并深刻影响着我对待工作和生活的态度，激励我在今后的岁月中脚踏实地的奋勇前行！

衷心感谢课题组长张勇东研究员，是他为我们营造了良好的科研环境，让我感觉进入了学习与科研的殿堂，使我们能够在和谐互助的氛围下潜心科研。感谢张老师对我各方面的支持和帮助，并对我的科研工作给了很多重要的指导和建议，使我能够快速地提升自己的综合能力，能在张老师领导的科研团队中学习工作是我宝贵的人生经历。

衷心感谢微软亚洲研究院的梅涛研究员，感谢他的支持与信任，让我有机会在国际化的科研环境中学习成长。他的指导与帮助，开拓了我的科研思路，训练了我独立进行科研的能力。他严谨务实、精益求精的科研作风让我非常钦佩，受益匪浅。在微软亚洲研究院实习是我非常宝贵的人生经历。

衷心感谢组长唐胜副研究员对我研究工作的支持及建设性的指导与帮助，唐老师深厚的学术功底、实事求是的探索精神给我了莫大的启迪和鼓舞，他对自己的高标准严要求是我学习的榜样。感谢我的第一任组长夏添老师，是他引导我进入美妙的科研世界，并在学习和生活中给了我非常多的指导与建议，使我收获了很多重要的指导和启发。本文的研究工作离不开两位小组长无私的帮助。

衷心感谢曹娟副研究员、张冬明副研究员、代锋副研究员、高科副研究员、顾晓光老师、郭俊波老师、崔洪亮老师、马宜科老师，感谢他们对我学习、工作和生活上的无私指导和帮助，几位老师高标准、严要求的治学态度帮我打下了坚实的基础，他们对我的关怀、鼓励和指导使我能够在研究的道路上不断前进。

衷心感谢行政办公室的刘玉东老师、任菲老师、刘卫玲老师和高姗姗老师，他们细致、辛勤的工作为我的生活、学习和工作提供了极大的便利和帮助。

衷心感谢研究生部的李琳老师、宋守礼老师、卢文平老师、冯钢老师、张平老师、李丹老师、周世佳老师和郭晓康老师，感谢他们在日常生活和学习中给我的关怀和帮助。

衷心感谢已经毕业的黄磊、陈智能、颜成刚、冯柏岚、张伟、谢洪涛、宋一丞、佟玲玲、庞琳、吴潇、汪文英、张旭、华秀峰、宋砚、苞蕾、庄东晔、张雷刚、刘毅、毛震东、宋砚、俞力克等师兄师姐对我的热情帮助，感谢褚令洋、张磊、张俊、杨晓鹏、高兴宇、吴波、王刚、王学辉、夏俊海、杨德杰、欧阳哲、李家宏、张曦珊等同学在工作和学习中的帮助，特别感谢同一个小组的万吉、董绍辉、张亮、王宇、梁飞蝶、韩淇、姜幽丞、王旭、曹阳、徐作新、张雅琳、高宝强、王学明、袁艳、陈慧、王斌、张蕊、李灵慧等同学，与他们一起做课题给我的研究生生活留下了无数美好的回忆。

衷心感谢美国罗切斯特大学的罗杰波教授、刘济教授，微软亚洲研究院的姚霆博士，感谢你们在学术之路上对我的提携和帮助。感谢在微软亚洲研究院一起工作学习的吴岳、刘恒、汪洋、杨飞彬、潘瀛炜、唐傲、于俊杰、杨绪勇、沈旭、付建龙、徐俊等人，感谢美国罗切斯特大学的游权增、李运成、胡天冉、杨希桐、仲浩辰、宋林峰、袁建博、KT、庞然、刘齐光等人，与你们相识我感到非常荣幸。

感谢我的父母，他们的理解、支持与鼓励使我有勇气面对碰到的每一个困难，他们全心付出才使得我有时间和精力完成学业。感谢我的爷爷、奶奶、姥爷、姥姥、舅舅、舅妈等亲人对我的关心和照顾。感谢我的未婚妻古晓艳，感谢她始终如一的理解、支持、包容与爱，感谢她对我的鼓舞、帮助以及她为我所付出的一切。

最后，衷心感谢百忙之中抽出时间审阅我的论文的各位专家！衷心感谢所有曾经关心和帮助过我的人们！

刘武

2015 年 5 月

作者简历

【基本信息】

姓名：刘武　　性别： 男　　出生日期：1987.12.01　　籍贯：山东泰安

【学习经历】

2009 年 09 月-2015 年 07 月	中国科学院计算技术研究所	硕博连读
2005 年 09 月-2009 年 07 月	山东大学	本科

【攻读博士学位期间发表的论文】

[1] **Wu Liu,** Tao Mei, and Yongdong Zhang, “Instant Mobile Video Search with Lay ered Audio-Video Indexing and Progressive Transmission”, **IEEE Transactions on Multimedia**, vol.16, no.8, pp.2242-2255, 2014, (**SCI 检索，影响因子: 1.776**).

[2] **Wu Liu,** Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li, "Accurate Estimation of Human Body Orientation From RGB-D Sensors", **IEEE Transactions on Cybernetics**, vol.43, no.5, pp.1442-1452, 2013 (**SCI 检索，影响因子: 3.781**);

[3] **Wu Liu**, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo, “Multi-Task Deep Visual-Semantic Embedding for Video Thumbnail Selection”, IEEE International Conference on Computer Vision and Pattern Recognition (**CVPR**), 2015, (**CCF A 类会议，EI 检索，**已录用)

[4] **Wu Liu,** Tao Mei, Yongdong Zhang, Jintao Li, and Shipeng Li, “Listen, Look, and Gotcha: Instant Video Search with Mobile Phones by Layered Audio-Video Indexing”，**ACM Multimedia**, 2013, pp. 887-896, (**CCF A 类会议，Oral Presentation**，EI 检索).

[5] **Wu Liu,** Feibin Yang, Yongdong Zhang, Qinghua Huang, and Tao Mei, “LAVES: An Instant Mobile Video Search System Based on Layered Audio-Video Indexing”, **ACM Multimedia**, 2013, pp. 409-410. (**CCF A 类会议，**EI 检索)

[6] **Wu Liu,** Tian Xia, Ji Wan, Yongdong Zhang, and Jintao Li, “RGB-D based Multi-Attribute People Search in Intelligent Visual Surveillance”, International Conference on Multimedia Model (**MMM**), 2012, pp. 750-760. (**Oral Presentation**，EI 检索)

[7] Yicheng Song, Yong-Dong Zhang, Juan Cao, Tian Xia, **Wu Liu**, and Jin-Tao Li, “Web Video Geolocation by Geotagged Social Resources”, **IEEE Transactions on**

Multimedia, vol.14, no.2, pp. 456-470, 2012, (**SCI 检索**, 影响因子: 1.776)

【攻读博士学位期间专利申请情况】

[1] 专利：夏添，**刘武**，万吉，张勇东，李锦涛，一种检测运动物体相互靠近和/或接触的方法和系统,（**已授权**，专利号：201110254404.X）

[2] 专利：Tao Mei, Shipeng Li and **Wu Liu**, “Mobile Video Search”，已申请

【攻读博士学位期间参加的科研项目】

- 国家 973 计划项目（2007CB311105）
- 国家 863 计划项目（2014AA015202）
- “十二五”国家科技支撑计划重点项目 （2012BAH39B02）
- 国家 242 信息安全计划项目（2009A53）
- 国家自然科学基金项目（61173054，61100087）
- 北京市自然基金面上项目 （4152050）
- 三星通信研究院合作项目 （4920125200）
- 华为中央研究院合作项目 （YB2013080040）

【攻读博士学位期间的获奖情况】

- 2014、2012 年 两次荣获 国家奖学金
- 2013 年 荣获 中科院计算所“夏培肃”奖学金
- 2014、2012 年 两次荣获 中国科学院大学三好学生标兵
- 2011、2010 年 两次荣获 中国科学院大学优秀学生干部
- 2010 年 荣获 中国科学院大学优秀党员
- 2013 年 荣获 中科院计算所优秀志愿者
- 2013 年 荣获 百度 Hackathon 编程开发大赛第一名
- 2012 年 荣获 中科院计算所第三届技术创新大赛二等奖
- 2013 年 荣获 中国科学院移动互联网应用青年创意大赛二等奖
- 2012 年 荣获 盛大云计算创意&开发大赛最佳创意奖
- 2013、2011、2010 年 三次荣获 中国科学院大学三好学生