



清华高性能计算案例通讯

(2007 年第 1 期, 创刊号)

清华信息科学与技术国家实验室(筹)公共平台与技术部编

2007 年 3 月 12 日

[编者按]

在理论研究、试验科学这两个基本研究手段之外, 计算科学正在成为基础科学研究的“第三维”。微软公司 2005 年 7 月发布了 40 余位国际上有影响的科学家历经 4 个月讨论形成的战略性咨询报告——《2020 年科学前瞻》, 列举了 21 世纪面临的七个全球性科学挑战, 即地球生命支持系统、生物学(细胞、免疫系统、大脑)、全球流行病学(SARS, 禽流感、口蹄疫、艾滋病等)、医药学、宇宙学、生命起源、未来能源, 指出这些问题的解决无不依赖于超级计算平台以及构筑于其上的对 TB 乃至 PB 级海量数据存储与分析处理的能力。

清华信息科学技术国家实验室(筹)建设的高性能计算平台, 已于 2005 年 11 月开放运行。该平台拥有清华大学性能最高的集群计算机系统, 采用高速 64 位处理器, 共 128 个结点、256 个 CPU, 峰值速度每秒 1.3 万亿次, 其即将扩充至 160TB 以上的存储能力居全国高校及研究单位之前列, 可为大规模计算提供坚实的保障。系统安装了各类编译工具(C, C++, Fortran, MPI)及相关数学库(如 MKL、SCALAPACK、PETSc 等), 且配备了科学计算可视化系统。目前已完成的作业数超过 1 万个, 涵盖了物理、化学、应用数学、材料、力学、电子、自动化、计算机、核技术、航天航空、生物信息、石油、电机等众多学科领域。

2006 年 9 月, “清华高性能计算应用专家委员会”成立, 顾秉林校长为此题词“祝清华高性能计算应用专家委员会成立! 加强高性能计算应用推广工作, 为清华跻身世界一流大学作贡献”。这标志着该平台的开放运行进入了新的阶段。目前它已纳入《清华大学实验室开放基金》, 对经用户专家委员会批准的校内用户, 可享有一定比例的机时费补贴。

我们诚挚地欢迎对高性能计算有强烈需求的广大用户来充分地利用这个平台, 尤其是那些冲击国际前沿水平、涉及重大基础理论研究或国民经济重大应用的课题。

我们将陆续编发在该平台上成功运行的典型案例, 为广大潜在用户推开一扇了解高性能计算之窗, 并最终打开进入高性能计算之门, 以更好地促进高性能计算的推广应用。

我们的服务理念与承诺是:

**以专业的高性能计算服务,
促一流的高水平科研成果!**

[案例]

视频检索评测（TRECVID）2006 之高层特征抽取

【研究单位】计算机科学与技术系

【用户】李建民博士（助理研究员）

【课题来源】973 项目（2004CB318108）

【科学背景及研究意义】

1. 高层特征抽取

随着网络中视频内容的日益增长，对视频进行基于内容的检索的需求日益迫切。国内外许多大学和研究机构纷纷开展视频检索的研究，开发出一些原型系统，有些已经投入商业运行，如 IBM 的 marvel，阿姆斯特丹大学的 mediamill，GoogleVideo，Blinkx 等。

视频的内在语义与它的表现形式（颜色、纹理、形状等底层特征）之间存在非常复杂的多对多关系。目前，计算机多是在低层次的特征空间内处理视频的，而用户的需求往往是处于高层次的语义空间中，二者差异巨大。这就是众所周知的语义鸿沟 (Semantic gap) 问题，导致低下的检索性能。

克服语义鸿沟的一个方法就是在底层特征和用户需求之间增加一个语义概念层，定义一些基本的语义概念，比如场景类概念（室内/室外、水景、城市等）、物体类概念（如汽车、建筑、动物等）等。然后分别建立从底层特征到概念及从概念到用户需求的两层映射，从而将所谓的语义鸿沟一分为二。用户可以直接使用概念来描述自己的需求，因而更为直接和方便。近几年基于语义概念的视频检索方法受到研究人员的广泛重视，而高层特征抽取就成为其中的一个研究重点。

所谓高层特征抽取（即高层概念检测或语义建模），就是构建从底层特征到语义概念的映射关系——语义概念的模型，对镜头内容自动标注预先定义的概念。例如，对于图 1 这个关键帧（一个镜头的代表帧），需要自动给出人物、脸、警察、走路或跑步、道路、户外等等标记。高层概念检测还不是真正的图像理解，只是标注某个概念是否出现，而并不作更深入的分析，比如物体出现的位置，人物识别等。虽然如此，借助这些简单的语义信息，可以在较高层次上描述视频的内容，从而为克服语义鸿沟开辟一个道路，一定程度上提高检索的性能。



图 1 一个关键帧示例

2. 视频检索评测 (TRECVID)

视频检索综合了图像处理、语音处理、文本处理、模式识别、机器学习等多种技术。研究者已经从特征表示、分类方法等多个方面提出了许多方法。然而在多数情况下，人们还无法从理论上证明某种方法的优越性，只能通过细致的实验评价它的有效性。因此，需要公共的数据集和评价方法以便比较各种方法。

TREC Video Retrieval Evaluation (TRECVID)正是在这种背景下诞生的，其目的是为了通过公开、可度量的评测来促进针对基于内容的视频检索的研究。TRECVID 由美国 Advanced Research and Development Agency (ARDA) 和 National Institute of Standards and Technology (NIST) 资助。自 2001 年以来，TRECVID 每年举行一次，其影响力日渐扩大，参加的单位逐渐增加，一些相关文章得到广泛的引用，已经成为视频检索领域中的国际性权威评测。

TRECVID 一般每年设三到五个子任务，其中可能有一个探索性的任务不进行评测。这些任务涵盖了视频检索的主要方面，比如结构分析（镜头边界检测等）和语义分析（低层特征抽取和高层特征抽取等），以及如何利用这些信息进行有效的检索（自动、手工、交互模式的搜索等）。

TRECVID 以真实应用为目标，复杂程度也日益增大。在 2005 年，视频数据已经达到大约 170 小时，英语、汉语和阿拉伯语三种语言，六个电视台；高层特征（语义概念）有 39 个，评测其中预先指定的 10 个。而在 2006 年，视频数据已经达到大约 330 小时，英语、汉语和阿拉伯语三种语言，八个电视台的十一个节目；高层特征（语义概念）仍为 39 个，但是评测其中的 20 个，且不预先指定。因此，TRECVID 的难度很大。

【国内外相关工作】

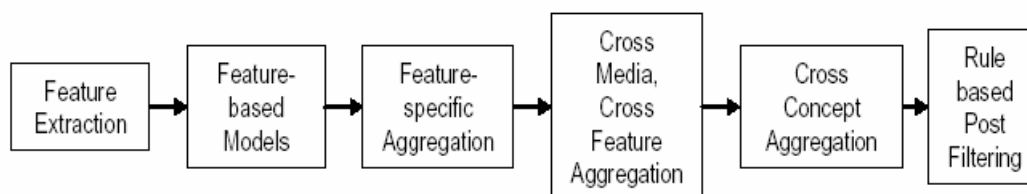


图 2 高层特征抽取的通用框架 [Naphade, acm mm2004]

由于高层特征抽取的重要性和 TRECVID 的推动，国内外许多机构都展开了这方面的研究工作。从图 2 中可以看出，高层特征抽取的第一步是底层特征抽取（各种层次的视觉特征、声音特征、文本特征、Metadata 等）。第二步是 Feature-based Model 的学习。第三步是信息融合，综合各种特征、模态甚至概念等各种信息，产生较为可靠的结果。最后进行一些后处理以进一步提高检测的准确性。

在高层特征抽取方面，IBM Watson 研究中心一直走在前列。在 2005 年，被评测的 10 个概念的 mean average precision 大约是 0.35 左右，是当年的最好结果。其他较好的单位有哥伦比亚大学，卡耐基梅隆大学，新加坡国立大学等。2005 年各个参加单位结果的中间值为 0.16 左右。

【本课题成果】

在过去的工作基础上，2006 年，本课题组继续参加了 TRECVID 2006。在高层特征抽取子任务中，以全部 20 个概念的 mean inferred average precision 为评价指标，本课题组取得的最好结果（0.192）列第一，且在 top 5 中占前四位（IBM 0.177 列第五）。另外以 20 个概念单独的 inferred average precision 为评价指标，本课题组的最好结

果有 6 个概念位于第一，2 个概念位于第二，4 个概念位于第三。参见图 3。

2006 年共有 IBM, CMU, Columbia, Oxford 等 28 个单位完成了这个子任务，共提交 125 组结果。国内有清华大学、复旦大学、浙江大学、微软亚洲研究院、香港城市大学提交。本研究工作得到 973 项目（数字内容理解的理论与方法之基于内容的多媒体信息检索）的资助，同时得到国家实验室公共平台与技术部和 Intel 中国研究中心的大力支持。

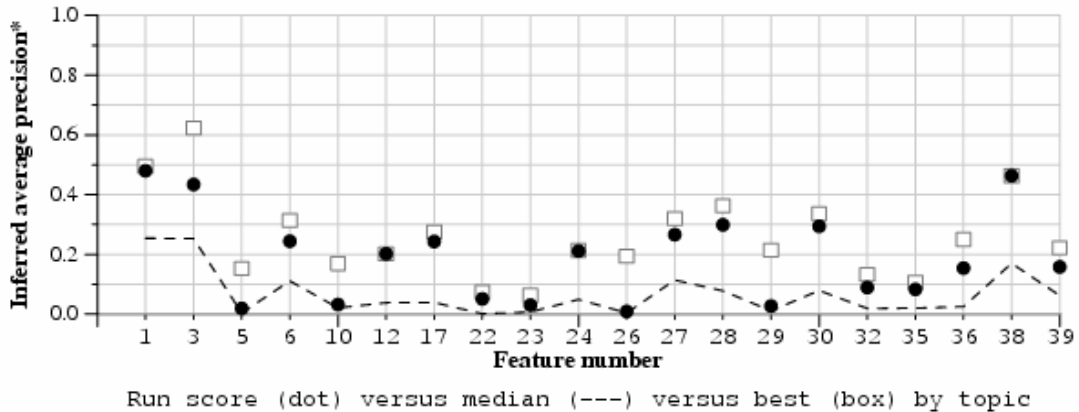


图 3 本课题组取得的最好的 run 性能，纵轴为 inferred average precision (评价指标)，横轴为概念的标号。黑点是本课题组的 run 中每个概念的 inf_ap，方框为所有单位提交的 run 中最好的 inf_ap，虚线为中间值。所谓一个 run，就是提交的一组针对所有概念的检测结果。

【高性能计算平台利用情况】

高层特征抽取是一个非常需要计算资源的任务。NIST 要求提交 39 个概念在 8 万多关键帧上的检测结果。为了进行检测，必须进行提取特征，构建模型及测试共三部分。其中构建模型最花时间，由于本研究采用了 SVM 方法，因此需要解许多的大规模的二次规划的问题。

针对这 39 个概念，本课题组在 2006 年暑假期间总共作了 39×110 个 SVM 分类模型，而且每个分类模型需要利用交叉验证对参数进行 grid 方式的选择。每个模型涉及到 7 万多的样本，因为 SVM 训练过程的计算复杂度大于 n^2 ，对这些模型的训练占了测评过程约 80% 的时间。另外有很多的 kernel 类型涉及到计算 emd 距离，更加剧了计算的复杂度，因为这部分的数量级增长和区域的个数成立方的比例，因此 13 个区域和 10 个区域的特征数据分别用到了相对而言 2000 和 1000 倍的计算资源。同时数据的吞吐也是非常大的，都在数百 M 的级别，仅训练好的模型就占了数十 G 的磁盘空间。因此在运算的时候，必须及时挪走计算完的模型，以便给新的数据和模型留地方，以免磁盘配额超限。

这是一个非常适合并行处理的任务。因为各个子任务之间没有耦合，并行计算的粒度自然就可以划分的很粗，所以 $39 \times 110 = 4290$ 个模型可以完全并行计算。本课题组当时在重写 matlab \rightarrow c 代码的情况下使用了国家实验室公共平台与技术部集群计算机系统 60 个计算节点，对其中比较耗时的涉及到 emd 距离的 kernel 类型，提交作业 3 天之后，2/3 部分的任务就完成了。如果没有这个计算平台的强有力支持，将无法赶在结果提交的期限之前及时完成任务。

编辑：林皎 审核：孙茂松 联系电话：010-62798983 E-mail: linjiao@tsinghua.edu.cn

http://www.tnlist.org.cn/pages/dept_gonggongpingtaiyujishu.jsp