JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

# 中国图象图形学报

2013
11
VOL.18
ISSN1006-8961
CN11-3758/TB


图像色彩视觉研究 P1536

# 中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）




第18卷第11期（总第211期）
2013年11月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊

**主管单位** 中国科学院
**主办单位** 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

**主　　编** 李小文
**编辑出版** 《中国图象图形学报》编辑出版委员会
**邮政信箱** 北京9718信箱
**邮　　编** 100101
**电子信箱** jig@irsa.ac.cn
**电　　话** 010-64807995 010-82614429
**网　　址** www.cjig.cn

**广告经营许可证** 京朝工商广字第0361号
**总 发 行** 北京报刊发行局
**订　　购** 全国各地邮局
**海外发行** 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
**印刷装订** 北京科信印刷有限公司

**Journal of Image and Graphics**

Title inscription: Song Jian | Monthly, Started in 1996

**Superintended by** Chinese Academy of Sciences
**Sponsored by** Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

**Editor-in-Chief** LI Xiaowen
**Editor, Publisher** Editorial and Publishing Board of Journal of Image and Graphics
**P.O.Box** 9718, Bejing, P.R.China
**Zip code** 100101
**E-mail** jig@irsa.ac.cn
**Telephone** 010-64807995 010-82614429
**Website** www.cjig.cn

**Distributed by** Beijing Bureau for Distribution of Newspapers and Journals
**Domestic** All Local Post Offices in China
**Overseas** China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
**Printed by** Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

**国外发行代号** M1406
**国内邮发代号** 82-831
**国内定价** 50.00元

2013年11月 第18卷 第11期（总第211期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

# 目次

融合对象性和视觉显著度的单目图像2D转3D(第1478页)

网格分割的3维网格模型非盲水印算法(第1529页)

基于参考图像的色粉画模拟(第1542页)

Volume 18, Number 11
Published November 16, 2013

# CONTENTS
## JOURNAL OF IMAGE AND GRAPHICS

Single-view image 2D-to-3D conversion based on objectness and visual saliency(P1478)

Non-blind watermarking algorithm for 3D mesh models based on mesh segmentation(P1529)

Color transfer with multiple parameters by combining the scaling and mean values(P1542)

E-mail: jig@irsa.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS


**中图法分类号**：TP301.6　　**文献标识码**：A　　**文章编号**：1006-8961(2013)11-1383-15

**论文引用格式**：中国计算机学会多媒体专业委员会. 多媒体技术研究:2012——多媒体数据索引与检索技术研究进展[J]. 中国图象图形学报,2013,18(11):1383-1397. [DOI:10.11834/jig.20131101]

# 多媒体技术研究:2012
## ——多媒体数据索引与检索技术研究进展

中国计算机学会多媒体专业委员会

**摘　要**：随着计算机网络、社交媒体、数字电视和多媒体获取设备的快速发展,多媒体数据的生成、处理和获取变得越来越方便,多媒体应用日益广泛,数据量呈现出爆炸性的增长,已经成为大数据时代的主要数据对象。然而由于多媒体数据本身的非结构化特性,使得多媒体数据的处理和检索相对困难。如何有效地存储、组织和管理这些数据,如何有效地按照多媒体的内容和特性去存取和检索这些数据,已经成为一种迫切的需求。面向大数据时代多媒体数据的索引和检索问题,围绕视频事件检测和标注、高维索引结构、跨媒体搜索3个研究方向,系统阐述了2012年度的技术发展状况,并对未来的发展趋势进行展望。

**关键词**：多媒体;事件检测;索引;跨媒体;检索

# Researches on multimedia technology, 2012
## ——Multimedia data indexing and retrieval technology: a review

CCF Multimedia Technology Committee

**Abstract**: With the rapid development of computer network, social media, digital television and communications technology, generating, processing and access to multimedia data has become more convenient. Multimedia applications are increasingly widespread, and the amounts of multimedia data are showing the explosive growth. In the era of big data, multimedia data has become a major data objects. However, due to its unstructured nature of multimedia data, the processing and retrieval of multimedia data is relatively difficult. How to effectively store, organization and management of these data, how to access and retrieve data in accordance with the multimedia content and features, has become an urgent need. This report is organized around video event detection and annotation, high-dimensional index structure and cross-media search three research directions, and elaborates their development status in 2012 and future development trend outlook.

**Key words**: multimedia; event detection; index; cross-media; retrieval

## 0　引　言

当前,全球网络的发展和普及已经达到空前的规模,网络已成为人们日常生活环境中的重要部分。来自微博、手机、社交网站、新闻网站以及多媒体共享网站中的文本、图像、视频和空间属性等多媒体数据每天都在以惊人的速度迅猛增长,网络多媒体数据与现实社会之间相互作用的范围之广和程度之深史无前例,在社会、经济和文化等诸多方面发挥着巨大效用,影响着无数人的思维、表达和行为方式。与文本数据相比,多媒体数据提供了大量有用的信息,其内容更加丰富、直观和生动。一方面,丰富的多媒体数据包含的海量信息,这是其他媒体类型所无法

**收稿日期**:2013-09-05;**修回日期**:2013-09-20

**基金项目**:国家自然科学基金项目(61225009,61202300,61173114);国家重点基础研究发展计划(973)基金项目(2012CB316400)

**通讯作者**:徐常胜,E-mail:csxu@nlpr.ia.ac.cn;于俊清,E-mail:yjqing@hust.edu.cn;吴飞,E-mail:wufei@cs.zju.edu.cn

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 11, Nov. 2013

比拟的;但在另一方面,其日益庞大的数据量、非结构化的数据形式和内容的多义性,又为方便用户交互操作设置了障碍,影响了它发挥更好的作用。人们已经习惯于在互联网上查找各种信息,目前,很多搜索引擎已经能很好地解决了文本的搜索问题(如Google、Baidu 和 Yahoo 等),但是对图像和视频数据的检索仍无有效的方法,其主要原因是由于数据量大以及内容复杂的特点导致缺乏有效的手段对其建立适合检索的索引。要对这些数据进行高效的操作,如浏览、访问和缩略等,需要对多媒体内容进行基于内容的分析和合适的索引。

为了克服文本检索技术的局限性,基于内容的图像和视频检索在 20 世纪 90 年代应运而生。不同于文本检索使用图像的文本描述信息,基于内容的图像和检索利用图像和视频的视觉特征或其他语义特征来构建索引结构。基于内容的图像和视频检索一般分为 3 个部分:特征提取、高维索引构造和检索系统设计。本文将对 2012 年度的多媒体信息的索引和检索的发展情况进行综述性介绍,内容涉视频事件识别与标注、高维数据索引和跨媒体搜索。

## 1 视频事件识别与标注

### 1.1 国内外研究现状

视频事件识别与标注就是根据视频所体现的内容,通过提取、存储、分析视频中的运动特征,来分析视频中的事件,从而按概念对其赋予标号。视频事件识别与标注是建立视频索引,进而实现高效视频数据处理(包括缩略,浏览和检索等)的必要基础。

视频中事件识别是计算机视觉领域中备受关注的前沿方向,它是通过计算机对图像序列中的内容进行语义描述,自动地分析理解以知道场景中在哪里发生了什么,并以此来模拟人对视觉信息处理的全过程。目前,视频标注的研究主要集中在如何利用视频的特点同时结合机器学习理论的相关知识来提高标注的准确性。从机器学习的角度看,语义标注通常可以看成是一个监督学习的过程。首先将视频分割成片段,然后从中提取低层特征,针对各种低层特征的训练集建立模型,实现对未知片段的标注。对视频进行语义标注,可以利用机器学习的理论结合视频本身的特点,尽量减少人工标注过程,实现半人工或者全自动的标注,最终达到可实际应用的目标。在统计机器学习理论中,为保证模型的泛化能力往往需要大量有标记的训练样本对模型进行训练。对视频语义标注的问题中面临着训练样本不足的情况,利用大量未标注的样本进行训练分类的半监督学习方法成为了视频标注研究的热点。随着多媒体视频内容在广播和互联网爆炸性地增长,视频事件识别与标注具有高度的学术价值和广泛的应用前景。其研究成果已经应用于智能监控、高级用户接口、虚拟现实、运动分析、视频管理和检索等生产、生活领域,并逐渐展现出更为广阔的应用前景。

目前,在国内外的很多高校和科研机构都开展了视频事件识别与标注方面的研究。例如,法国国家信息与自动化研究院、英国雷丁大学、美国的麻省理工学院,加州大学伯克利分校,英国的牛津大学,剑桥大学等。另外,当前国际上一些权威期刊如 IJCV、PAMI 和重要的学术会议如 ICCV、CVPR、ECCV、ACM MM 等将视频事件识别与标注研究作为主题内容之一,为该领域的研究人员提供了更多的交流机会。在中国,很多高校和研究机构近几年才开展此项研究。比如中科院自动化所模式识别国家重点实验室、北京大学视觉与听觉信息处理国家重点实验室、微软亚洲研究院视觉小组和清华大学等。另外,当前国内一些权威期刊如《计算机学报》《软件学报》《中国图象图形学报》《自动化学报》等将视频事件识别与标注研究作为研究主题并为该领域的研究人员提供了交流的平台。

视频事件识别与标注的关键问题是如何有效地抽取和表述事件的特征,以及如何根据得到的特征进行建模和分类。近些年来,大量研究人员对以上问题进行了深入的研究,并取得了一定的成果[1-8]。视频中的行为事件是一个时变信号,事件识别与标注问题就是一个时变信号的分类问题,按照不同的标准分成不同的类型,如:基于产生式模型的方法和基于判别式模型的方法,模板匹配方法和状态空间方法,基于监督学习、无监督学习和半监督学习的方法等。下面将对国内外视频事件识别与标注的相关工作进行分析总结,对不同的方法进行对比分析。

#### 1.1.1 基于模板匹配的分析方法

模板匹配方法[9-17]可以分为帧对帧匹配方法和融合匹配方法。例如,Bobick[9]将行为图像序列表示为运动能量图和运动历史图,然后提取矩特征作为行为模板。Wang 等人[14]采用变形模板匹配的方法计算不同姿态间的距离。Babu[18]提出了一种利用运动历史图和运动光流历史特征进行识别的方

法。Carlsson 和 Sullivan [19]利用每类代表性的例子与视频的每帧匹配来识别网球中的正手和反手等姿态。Dedeoglu 等人[15]利用基于关键帧的方法建立了一个实时的系统来识别各种行为。Daniel 和 Edmond[16]采取基于例子的方法将长度不同的视频转换成一个固定长度的描述,基于这个描述训练分类器。Laptev 和 Lindeberg[20]引入了 Gallilean 变化得到了速度变化自适应的时空兴趣点,解决了不同相对运动速度下的基于时空兴趣点的事件行为匹配问题。基于模板匹配的方法通常计算量比较小,但对于时空域的变化比较敏感,鲁棒性差。

1.1.2 基于状态空间的分析方法

基于状态空间模型的方法[21-25]定义每个静态姿势作为一个状态,这些状态之间通过某种概率联系起来。任何运动序列可以看做是这些静态姿势的不同状态之间的一次遍历过程,在这些遍历期间计算联合概率,其最大值被选择作为分类事件行为的标准。这类方法主要以隐马尔可夫模型和动态贝叶斯网络为基本手段进行事件分析与理解。例如,Yamato 等人[23]通过引入隐马尔可夫模型识别网球运动中的基本动作。Starner 和 Pentland[24]提出了一种基于 HMMs 的系统来识别美式手语。Luo 等人[26]将动态贝叶斯网络引入事件识别。Dbn 用条件概率来形成联合概率,训练相对要简单;但是 Dbn 的设计要比 HMM 复杂得多。总之,基于状态空间的事件识别算法可以避免对时间间隔建模的问题,但是需要的训练样本大,计算复杂。

1.1.3 基于参数模型的分析方法

基于参数模型的事件分析算法是通过统计不同事件的概率分布,以包容现实中事件上的某些模糊变化。该方法包括产生式模型和判别式模型。产生式模型[26-29]是一种基于概率的不确定性推理模型,是公认的学习时间序列数据效果很好的模型。其思想是通过一个联合概率函数模拟随机观测数据和隐藏状态之间的关系。在事件识别过程,新的观测数据,根据已经获取的多个事件模型参数计算似然,比较取其大者为识别结果。用于事件识别的展开式模型多为有向图,最常见的就是混合高斯模型、马尔可夫随机场[27]、动态贝叶斯网络[26]、隐马尔可夫模型[28]、概率随机方法[29]等。判别式模型与产生式模型不同,产生式模型存在着严格的独立性假设,每个观测值被作为独立的单元对待,与序列中所有其他观察值无关,且认为当前观察值是由当前状态生成的,同其他状态相互独立,这样就无法表现观察值序列中存在的长距离依赖关系。与之不同的是,判别式模型直接对条件概率进行建模,它认为观察值决定状态,克服了产生式模型的严格独立性假设,对于特征数据之间的相互关系没有任何的限制,能融合各种特征到模型中。在事件识别领域,广泛应用的判别式模型包括支持向量机、条件随机场和神经网络[30-32]。

1.1.4 基于半监督学习的分析方法

许多现有的方法[33-36]是基于监督学习的方式,为了得到好的识别效果,这些方法需要大量的标签样本训练事件模型[36]。但是,有标签的样本通常需要大量的人工标注,因此得到这些样本是比较困难和昂贵的。因此,近几年来,许多研究者将注意力转移到了基于无监督和半监督学习的事件分析方法的研究中。由于事件比较复杂,因此无监督的学习方法效果并不好,而半监督的学习方法是监督和无监督学习方法的折中,是目前事件分析的一个热点研究方向。基于半监督方法能够根据少量的标签样本和大量的无标签样本得到一个好的训练模型。由于这类方法能够自动地标记样本,更加适用于实际场合的事件识别与检测以及未知事件的发现。Manor 与 Irani 提出了一种自动建模的方法[37]。Niebles 等人提出了一种概率图方法[13]。但是,这里有很少的方法是基于半监督的,而这种方法能够同时利用有标签和无标签的样本。Zhang 等人提出了一种半监督建模方法[38],即首先通过基于监督学习的方法建立一种正常事件模型,通过迭代自适应的方法自动地建立异常事件的模型。

### 1.2 国内外研究进展比较

视频事件识别与标注是当前国内外非常重要的研究领域,拥有强大的应用推动力。在国内,从事视频事件识别与标注的单位主要有浙江大学、清华大学、中国科学院自动化研究所、微软亚洲研究院等。相关课题受到高度重视,得到了国家高技术研究发展计划、国家重点基础研究发展计划和国家自然科学基金等的资助。与前几年相比,理论研究和产业应用水平逐年提升,在国际学术刊物和重要国际会议上发表了越来越多的论文,申请了越来越多的国际专利,并且与国际同行的合作也更加紧密。在国外,如在美国,随着网络技术的发展和大量视频的产生,政府全力推进相关技术的发展,开展了大量关于视频事件识别与标注项目的研究。

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 11, Nov. 2013

2001 年,NIST(美国国家标准技术研究所)开始组织 TRECVID 评测,并且从 2002 年开始将视频标注列为一项单独的任务,这标志着大家对这个领域开始了系统性的研究。值得一提的是,国内很多高校和研究机构,如清华大学、复旦大学、中国科学技术大学以及微软亚洲研究院等,很早就开始关注这个领域,积极参加 TRECVID,并且在各项任务评比中取得了很好的结果。与国际研究水平相比较,国内的这些高校和研究机构毫不逊色,甚至于在很多技术上处于领先地位。尽管中国在视频事件识别与标注方面取得了一些成果,但与美国、法国、德国等国际先进国家相比,无论是在研究还是在产业应用上,还是有一定的差距的。

总地来说,视频识别与标注技术仍处于理论研究阶段。由于一般的识别方法需要处理庞大的数据量,计算代价非常高,而且,这些方法多数会极大地受到视频数据复杂性等因素的影响,因此,要做到灵活、实时的应用仍存在许多难以解决的问题。目前,视频事件识别与标注还是集中于视频中目标的跟踪、特征提取、简单事件识别等问题,虽然近年来利用机器学习、人工神经网络等工具构建视频理解的统计模型的研究有了一定的进展,但依据事件识别与标注进行视频的有效管理和理解仍旧处于初级阶段。而且,对于视频事件识别与标注来说,目前国际上没有通用的、权威的大规模的数据库,算法的评估与实验都是在各自小样本数据库上模拟进行的。数据库在构建时限定了许多约束条件,比如:静止的摄像机,相对比较简单的背景,简单的事件类别等。

### 1.3 发展趋势与展望

尽管视频中事件识别与标注技术正展示出广阔的应用前景,众多的科研人员也已经投入了大量的精力进行相关方面的研究并涌现了许多有效的算法,但是很多工作都是一些探索性的研究,它们仅限于对一些简单场景或严格限制的场景中简单的事件进行分析与理解,对于相对复杂的场景和行为,现有的研究成果还难以处理。目前各自算法的测试和评估都是在少量的几个数据库上完成的,缺乏普遍性和说服力,很多技术迫切需要完善。影响其进一步发展的主要原因在于以下几个方面:

1)事件表述　事件表述是事件分析的基础,因此很多研究工作着力于有效的特征提取和融合。但是,目前的特征提取和融合方法对于复杂场景下的事件分析仍然存在不足。所以,如何构造一种有效、描述准确度高并具有良好鲁棒性的事件表述也是事件分析研究中的难点之一。

2)事件模型　就事件模型而言,不论产生式模型还是判别式模型,目前的模型结构都是研究者根据先验知识事先定义好的,很少自动获取事件的结构,这难免对事件本质产生一定程度的束缚,不能够完全地,真实地表达事件的本质。我们知道模型越简单越有效,但是真实世界中的事件总是错综复杂的,到底哪一种模型才是有效的可以泛化的值得深入地研究探讨。

3)高层语义理解问题　目前对事件分析的研究还是停留在约束条件下简单事件的分析,其在实际场景中的应用还有很大距离。事件分析的最终目的是要电脑分析和理解人的行为,这种理解不仅要识别人的基本行为,还要能根据识别的结果做出反应。

4)鲁棒性、实时性和精度　事件描述与事件的语义存在鸿沟,因此需要找到合适的学习方法来鲁棒地解决在事件描述不好的情况下的事件识别与标注问题。另外,对于不同的应用场合,事件行为识别算法需要兼顾鲁棒性、实时性和精度等多方面因素。针对不同的应用场合,如何实现鲁棒性、实时性和高精度等因素的折中,也是事件分析技术中亟待解决的问题之一。

虽然目前基于机器学习的视频识别与标注已经取得了很大进展,但是其距离实用化仍有一定距离,要真正实现大规模数据集高效而精确的事件识别与标注,仍需进行进一步的研究。这里简单阐述几种可能的改进和扩展。

1)挖掘事件相关性的研究　视频片段间的事件是具有一定关系的,如果能够对这种关系加以挖掘,必然能够提高标注的准确性。然而目前的研究表明,通常是仅有一部分事件的标注结果会提高,但是也会降低另一部分事件的标注性能。如何设计更稳定的多事件相关学习算法,仍然值得进一步研究。

2)自适应特征提取　不同的事件概念需要通过不同的特征进行区分,为了对整个事件概念集合进行分析,需要尽可能多地提取各种特征参数,如颜色、边缘、纹理以及运动特征等。问题在于:往往仅有部分特征与待标注概念相关,冗余的部分不仅增加了计算负荷,而且可能会对最终标注产生负面的影响。因此需要研究针对特定的事件选择合适的特征参数的方法。这部分可以借助深度学习方法进行

改进。

3)鲁棒的相似性度量　目前的学习算法中采用的相似性度量,大部分都是直接根据样本间的距离定义的。然而,相似度不仅仅与距离有关,也应考虑到样本周围分布的相似性。

4)有效的模型训练方法　当前大多数方法为有监督的方法,即采用人工标记的方式来训练行为模型。这种方法需要人工标注获取各类行为的样本,因此将会浪费大量的人力。近年来,人们提出了自动获取样本,自动建立行为模型的方法。因此无监督或半监督的方法是重要的研究方向。

5)复杂场景下复杂事件识别与标注　目前许多工作主要是对简单场景下的简单事件进行识别和标注,并取得了较好的结果,而对于复杂场景下的事件识别和标注效率很低。现实生活中的很多场景都是比较复杂的,甚至是时常变化的,而且行为也不单单是一些基本的动作,因此复杂场景下复杂事件的识别和标注问题是一个值得进一步研究的问题。

## 2　高维索引结构

### 2.1　国内外研究现状

随着互联网和多媒体技术的快速发展,以图像为代表的多媒体信息呈爆炸性增长。面对海量的图像库,只有对图像进行有效地组织以便于浏览、搜索和检索,人们才能快速并准确地获取自己感兴趣和喜爱的图片。

基于本文的图像检索始于 20 世纪 70 年代末,数据库中图像首先用文本描述进行描述,然后构建一个基于本文的图像管理系统用于检索。当图像库的规模很大时(图像数量在百万级或者亿级以上),人工标注既费时又费力,其局限性越来越明显。此外,由于图像内容的丰富性以及图像文本描述本身带有强烈的主观色彩,因此,不同的人可能对同一幅图像有不同的理解,导致图像的人工描述存在差异。

为了克服文本检索技术的局限性,基于内容的图像检索在 20 世纪 90 年代应运而生。不同于文本检索使用图像的文本描述信息,基于内容的图像检索利用图像的视觉特征来构建索引结构。基于内容的图像检索一般分为 3 个部分:视觉特征提取、高维索引构造和检索系统设计。对于大规模图像库,对大量的图像视觉特征进行某种处理并直接存储在计算机内存以提高图像检索速度成为当前的研究趋势。更为关键的是设计一个高效的高维索引结构来组织图像的视觉内容,实现图像的准确和快速检索。

2.1.1　基于树结构的高维索引

KD 树虽然在低维空间中能有效降低搜索复杂度,但当特征向量维度超过 20 时,性能迅速下降,搜索复杂度接近于线性搜索,面临“维度灾难”问题。通过构造多个 KD 树并且同时检索这些 KD 树,Silpa等人[39]对 KD 树进行了改进并对包含 500 000 个 SIFT 的特征集构建索引。

Zhuang 等人[40]提出了基于对称编码的混合距离树(EHD-Tree)高维索引结构,将 k-NN 近邻搜索从高维空间转换到 1 维空间。

相对于用单一的特征描述视频或图像,用多种特征能更准确地描述它们。He 等人[41]提出了一种面向权值的多特征索引结构(MFI-Tree)。MFI-Tree 是度量空间的一种动态索引结构,通过与之匹配的 ADD-kNN 检索算法可以实现权值可变的多特征检索。

2.1.2　基于哈希的高维索引

基于哈希的高维索引分为两类:以 E2LSH 为代表的将特征点映射到低维欧氏空间,使用欧氏距离度量特征点之间的相似度;以谱哈希为代表的将特征点映射到低维汉明空间并保证欧氏空间中相近的特征点具有相似的二进制编码,使用汉明距离度量特征点之间的相似度。

1)位置敏感哈希

基于 p-stable 分布,Datar 等人提出了精确欧氏位置敏感哈希(E2LSH)[42]。针对欧氏空间的 c 近似最近邻问题,Andoni 等人[43]提出了近似优化哈希算法。文献[44]提出了两种查询扩展策略来改进 E2LSH:内部扩展和交互扩展。

为了降低 E2LSH 的存储空间需求,Panigrahy 提出了基于熵的 LSH[45]。在此基础上,Lv 等人提出了 multi-probe LSH[46]。为了提高查询准确率,multi-probe LSH 使用 step-wise 和 query-directed 两种探查序列来探查哈希表中多个邻近桶,使得查询对象的近邻特征点尽可能被检索到。

Jegou 等人提出了自适应 LSH 方法(QA-LSH)[47],为查询特征点从一组哈希函数中在线选择最合适的哈希函数。QA-LSH 虽然仅仅在哈希函数的构造方式上不同于 E2LSH,但是在查询速度和准确率上可以获得更好的平衡。利用 K-means 聚类方法构造哈希函数,Pauleve 等人提出了 KLSH

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 11, Nov. 2013

方法[48]。

2)哈希编码

哈希编码的思想是利用少量的二进制编码使得相似的特征点具有相同或者相似的二进制编码,通常使用汉明距离度量特征点之间的相似度。将RBM和Boosting相结合用于学习二进制编码,Torralba等人[49]实现了对百万级以上图像库进行编码。在此基础上,Weiss等人提出了谱哈希(SH)[50]。SH在编码效率上与Torralba的编码方法相同,但更易于实现。

在文献[51-52]的基础上,文献[53]提出了球形哈希(spherical hashing),将空间上位置邻近的特征点映射到相同或者相似的二进制编码。球形哈希使用超球体来划分特征空间,每个超球体与一个哈希函数相关联。

利用监督信息,监督式哈希索引方法可以获得比非监督哈希索引方法更好的检索准确率,例如,LDAH[54]和MLH[55]。然而,监督式哈希索引方法在训练学习阶段通常比非监督哈希索引方法花费更多的时间,不仅如此,当图像的标签数据太少或者噪声太多时,这类方法可能会出现过度拟合的现象。为了提升监督式哈希索引方法的性能,Wang等人[56]提出了半监督哈希(SSH)方法,在最大化标签数据和无标签数据哈希码的差异性和独立性的同时最小化经验误差。类似研究还包括基于核的监督哈希(KSH)[57]编码方法。

Shao等人[58]提出了稀疏谱哈希(sparse spectral hashing),将稀疏主成分分析(Sparse PCA)和增强相似敏感哈希(Boosting Similarity Sensitive Hashing, Boosting SSH)引入SH。超图谱哈希(HSH)[59]针对图像的社交内容设计哈希函数,实现大规模社交图像的快速检索。Zhou等人提出了一种标量量化方法[60],其根据每个特征向量的中位值进行量化和编码,无需任何训练过程。

2.1.3 基于视觉单词的倒排索引

词袋模型(BOF)作为首个基于视觉单词的倒排索引,引入自文本检索系统。BOF图像描述符之间的相似度通常用欧氏距离或者角度距离度量。

1)基于视觉单词的图像描述符

基于视觉单词的图像描述最初是由Sivic等人[61]人提出,利用图像的sift局部特征计算BOF图像描述符。BOF图像描述符向量的每个分量的权值根据tf-idf机制计算。

(1)构造视觉词汇

BOF图像描述符的高维度意味着需要训练大量的视觉单词。标准BOF[61]使用的原始K-means聚类方法在大规模视觉单词的训练和聚类上需要耗费大量时间,变得非常低效。层次K-mean[62]和近似K-means[63]方法可以用于代替原始K-means方法来提高视觉单词的训练和聚类的效率。

为了降低视觉词汇的存储空间和提高时间效率,文献[64]提出了随机位置敏感词汇(RLSV)。RLSV将LSH和随机森林相结合,通过一系列的随机线性二部划分来构造视觉词汇。Kuo等人[65]提出了无监督的辅助视觉单词(AVW)发现机制,提高了BOF的区分度。

(2)特征分配

传统的BOF描述符特征量化方法采用的硬分配存在视觉单词不确定性和视觉单词疑惑性两个缺陷。为了克服这两个缺陷,文献[66]提出了基于核的软分配,其首先计算特征点到每个视觉单词的权值,然后将特征点分配到权值大于预先设定阈值的视觉单词。文献[67]提出了描述符空间软分配将特征点分配到最近的$k$个视觉单词来减少量化误差。文献[68]提出了球形软分配将特征点自适应地分配到近邻视觉单词。

(3)其他基于视觉单词的图像描述符

基于BOF和Fisher核,Jegou等人[69]提出了局部聚合描述向量(VLAD)。VLAD比BOF具有更高的准确性但维度更低。将每个视觉单词对应聚类中最远特征点与聚类中心之间的距离作为阈值,Chen等人[70]通过移除位于两个视觉单词之间边界上的异常特征点,进一步提高了VLAD的准确性。

2)描述符压缩与编码

基于汉明嵌入(HE)和弱几何一致性(WGC),Jegou等人[71]提出了比BOF图像描述更加紧凑和准确的图像描述符。基于HE,文献[72]提出了一种近似BOF图像描述符miniBOF。一幅图像的标准BOF图像描述符通常需要10k字节左右存储空间,而当BOF描述符分解成16个miniBOF描述符时,一幅图像只需320字节。Jain等人提出了非对称汉明嵌入(AHE)机制[73],提高了HE的编码性能。AHE在没有增加任何开销的基础上提高了HE的编码效率。

描述符量化同样可以用于描述符的编码,文献[74]提出了积量化(PQ)将特征描述符向量分割为

m 个子向量并分别对其进行编码。与全局量化器相比,PQ 不仅使用更少的聚类中心还明显减少了存储空间。PQ 只是将描述符向量按照固定长度进行均分得到子特征向量,没有考虑到特征向量各个维度分量的实际分布。Brandt[75] 提出了将转换编码和 PQ 相结合的描述符量化方法。Chen 等人提出了残差向量量化(RVQ)[76] 方法,利用多个低复杂度的量化器来减少量化误差,使得对结构化描述符和非结构化描述符都能进行高效的量化。

3)基于视觉单词的倒排索引

基于视觉单词的倒排索引结构最初是在文献[61]中提出,每个视觉单词对应一个倒排列表,倒排列表中节点存储的信息是图像 ID 和视觉单词在该图像中发生的次数。

基于 PQ,文献[74]提出了基于非对称距离计算的倒排索引(IVFADC)用于特征描述符的最近邻搜索。基于 PQ,Babenko 等人[77] 利用多维表来构建倒排多维索引。Babenko 首先将特征描述符平分为 2 个子向量,然后分别在子向量所在特征空间上用 K-means 训练包含 $k$ 个聚类中心的量化器。将分别来自两个量化器的聚类中心组合为聚类中心对,形成一个 2 维表,从而构建了包含 $k \times k$ 个倒排列表的多维倒排索引结构。

文献[76]提出用 RVQ 的前 $L$ 层量化器构建倒排索引。如果每层量化器包含 $k$ 个聚类中心,那么就可以构造具有 $k^L$ 个倒排列表的倒排索引结构。

## 2.2 国内外研究进展比较

作为基于内容的大规模图像检索的关键之一,基于内容的高维索引结构是当前国内外这一研究领域的研究重点和热点,拥有广阔的应用前景。在国内,开展高维索引结构的研究单位主要有华中科技大学、浙江大学、清华大学、中国科学院计算技术研究所、微软亚洲研究院以及 IBM 中国研究院等,企业单位主要有百度、腾讯、360 以及搜狗等,并且这些单位都已推出了基于内容的图像搜索服务。近年来,随着国际合作越来越紧密,国内的研究者在国际重要学术期刊以及国际顶级学术会议上发表了越来越多的高水平论文,已与国外的研究处于同步阶段。

在国外,率先关注高维索引结构这一领域的研究单位主要有麻省理工学院、美国哥伦比亚大学、美国伊利诺伊大学以及牛津大学等,他们也各自研发出了一些基于内容的图像检索系统。企业单位包括微软以及谷歌等,并且这些单位都已在搜索引擎中推出了基于内容的图像搜索服务。虽然国外研究者们在高维索引结构的研究上比国内关注的更早,但随着高维索引结构成为国内的研究热点以及更紧密的国际合作,目前国内的研究水平和研究成果毫不逊于国外。近年来,著名的法国国家信息与自动化研究所(INRIA)在这一领域的研究非常活跃,开展了一些关于大规模图像索引与检索的研究项目:“ICOS-HD”对视频内容的高维描述符构建索引和进行可扩展性压缩,“Quaero”对多媒体(图像、视频等)和多语言文档的自动分析与分类,其中就包括对图像和视频检索的研究。

总体而言,国内外对构建高维索引结构的方法:基于树结构的高维索引、基于哈希的高维索引和基于视觉单词的倒排索引都投入了很大精力进行研究。相对于基于树结构的高维索引,基于哈希的高维索引和基于视觉单词的倒排索引受到国内外研究者更多的关注。在后两种高维索引方法的研究中,相对来讲,国内研究者更侧重于基于哈希的高维索引,尤其是哈希编码方法成为最近的热门研究方向。虽然目前国内外的研究取得了很多研究成果以及各大搜索引擎都推出了基于样例的图像搜索,但距离成熟阶段,尤其是针对海量图像的索引和检索,尚有一段距离,仍然存在许多问题亟待解决。在基于哈希的高维索引方面,其通过一系列的哈希函数,将欧氏空间中相近的描述符映射到汉明空间中相同或者相近的二进制编码。然而,二进制编码的汉明距离的相似度区分能力受限于编码长度,因此,在汉明距离有限区分度的前提下,如何设计哈希映射函数,进一步提高检索准确率值得进一步研究。在基于视觉单词的倒排索引方面,目前的倒排索引的构建过程是通过对一个样本集进行训练完成的。然而,互联网上的数据每时每刻都在快速增长,而已构建好的索引结构只能反映当时的数据分布,而不能实时地反映当前得到的数据。因此,如何设计对动态插入或者删除具有健壮性的倒排索引值得进一步探索。

## 2.3 发展趋势与展望

尽管目前的高维索引技术采用如特征降维或者描述符特征分解等方法来应对高维特征的“维度灾难”,但暂时还没有彻底解决该问题。因此,一个高效和健壮的高维索引技术仍然需要进一步研究。

基于视觉单词的倒排索引利用描述符压缩和编码,使得海量的特征描述符可以存入计算机内存。然而,目前这类索引方法通常使用的是图像低层特

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 11, Nov. 2013

征,而检索得到的低层特征距离相近的图像对用户来说,不一定是视觉感官上相似的。因而,如何将图像低层特征与一些语义信息相结合,构造更能准确描述图像内容的描述符值得进一步关注和研究。

目前在构建倒排索引时需要样本集训练聚类中心,而这些聚类中心只能反映当前数据集的分布情况。当插入一批新数据或者从数据集中删除一批数据时,由于数据分布已发生改变,因此已训练的聚类中心就不能反映当前的数据分布,进而影响检索准确率。针对这种问题,目前的索引方法采用重构索引来解决。但是,当数据集的规模很大时,重构索引需要花费大量的时间。因而,设计一个支持动态插入和删除的倒排索引仍然需要进一步的研究。

随着移动互联网和智能手机的出现与发展,人们之间的交流越来越依赖于智能手机。手机存储空间越来越大,此外,智能手机的处理器性能越来越强大,如今已有4核处理器甚至8核处理器。这意味着个人智能手机中的图像等多媒体信息也越来越多。因此,面对智能手机和移动互联网中大规模图像,如何使人们可以从其中快速地检索到相似并感兴趣的图像或许会成为一个新的研究热点。

## 3 跨媒体检索理论与技术

有别于传统的结构化和非结构化数据,网络上的数据是以文本、图像和视频等多种形式体现的,这些从不同渠道获取的文本、图像和视频等不同类型媒体及其与之相关的社会属性信息更加紧密地混合在一起,以一种新的形式,更为形象地表示综合性知识,反映个体和群体的社会行为,这种新的媒体表现形式称为“跨媒体(Cross-media)”。

跨媒体表现出4个基本属性:1)固有的自然属性。网页、图像、视频以及用户评价等结构化或非结构化信息具有异构、多阶和高维等自然属性;2)与现实生活密切相关的社会属性。网络平台上的跨媒体数据通过不同方式被赋予了评价、热度和偏好等社会属性,这些社会属性汇集成社会影响力,对跨媒体内容的语义理解产生主观影响;3)数据来源多样以及不同类型媒体数据相互关联和动态演化。来源于不同渠道的结构化与非结构化数据之间交叉关联,且跨媒体数据所蕴含的主题和事件通过一定机制传播并动态演化;4)丰富的表达和呈现能力。与人脑通过不同感官渠道对外界认知一样,通过跨媒体形式,能更为自然地展示客观世界及其包含知识。

下面将对跨媒体搜索中所涉及的跨媒体度量、跨媒体索引和跨媒体排序技术发展现状进行系统阐述。

### 3.1 国内外研究现状

#### 3.1.1 跨媒体度量

在跨媒体检索中,如何挖掘不同类型数据之间的内在联系,进而对跨媒体数据之间的相似度进行计算,是跨媒体检索要解决的重要内容。

在基于典型相关性分析(CCA)[78]的关联性建模过程中,由于从数据中提取的高维特征存在噪声和冗余,Sparse CCA[79]和Structured Sparse CCA[80]由此被提出,希望对高维特征进行选择之后再进行投影和计算数据之间的关联性,从而使得模型之间的关联性建模更加具有可解释性。为了对两种类型以上的数据之间关联进行建模,文献[81]将传统CCA拓展为多视图CCA算法,使其可以被用来对多种类型数据之间相关关系进行建模。泛化多视图分析(GMA)[82]拓展了CCA算法,GMA不仅在关联建模中可利用类别信息,还引入了核函数来反映数据的非线性分布。Zhuang等人提出了一种基于改进的稀疏典型相关性分析方法[83],并将其应用于跨媒体度量学习领域,其和文献[84]中提出的针对“原始—对偶”数据问题的Sparse CCA略有不同,该算法更为通用简洁,且易于实现。

一些通过主题建模来实现不同类型数据之间关联建模的方法也陆续被提出。Correspondence LDA(Corr-LDA)[85]是最早对LDA进行拓展来处理两种不同类型数据的方法,Corr-LDA对图像及其伴随标签文本之间的关联关系进行建模。但是在Corr-LDA这一方法中,图像中所蕴含“主题”是由其对应文本所蕴含“主题”直接得来的,这就限制了图像—文本关联关系的灵活性。文献[86]是对这一不足进行改进的一种方法,其认为图像中所蕴含“主题”可由文本所蕴含“主题”线性组合而来,这种方法比Corr-LDA灵活不少,但是仍然存在图像中所蕴含主题受到文本中所蕴含主题限制的问题。为了解决这一问题,文献[87]提出了多模态文档随机场(MDRF)的方法。MDRF采用马尔可夫随机场来刻画文档之间的关联关系。在MDRF中,一个文档可以由任意一种模态数据所构成,不同文档之间通过马尔可夫随机场相联系,区别在于不同模态字典间的差异。文献[88]提出了一种将“主题”分为公共

“主题”和私有“主题”的方法。该方法在对不同类型数据之间关联关系进行建模时具有更大灵活性。

通过字典学习来对不同类型数据之间关联性进行建模的手段也被陆续提出,如文献[84]应用字典学习来学习不同类型数据之间的关联关系。然而,全局共有系数空间的假设会导致不同模态数据在字典学习中重建系数被迫相同,从而限制了算法本身。因此,如何将字典学习更加自然应用于多模态数据,是一个热点问题。文献[89]提出了一种基于结构信息的监督耦合字典学习方法,将其应用于跨媒体检索。这一方法通过耦合字典学习来求取不同类型数据的字典及每一数据所对应的稀疏系数,同时求取一个映射函数来实现不同类型数据的映射。

互联网中异构数据间的关联通常是复杂高阶的。与传统的图表达不同,在超图(hypergraph)中,每一条超边(hypergraph)可以连接任意多个图中的顶点,从而将存在高阶关联的数据进行关联。因而,超图非常适合对跨媒体数据之间的关联进行建模,并在近年来成功地应用到了跨媒体检索中。在文献[90]中,Yu 等人通过引入超图表达将图像的属性信息与图像的底层视觉特征进行建模,并且应用到了图像相似性的检索中。在文献[91]中,Hong 等人提出了语义关联超图对图像及其关联进行建模。在文献[92]中,Hong 等人将超图应用在了多视图数据的图像块对齐工作中。通过将图像块看做超图中的超边并将相邻块的相似度超边权重,定义了多视图拉普拉斯矩阵并基于此实现了图像块对齐。

### 3.1.2 跨媒体索引

跨媒体索引是传统高维索引的扩展,它的基本思路是将不同模态的高维数据映射到一个统一表达的低维索引空间,使得不同模态内具有相近“语义”的数据在索引空间内具有相近或相同的表达。

跨媒体哈希索引问题在 CMSSH[93] 中首次被提出并给出如下解决思路:对于两个维度不同的不同类型数据,CMSSH 目标是分别学习一组组哈希函数,使得对于存在已知语义关联关系的两个不同类型数据,它们经过各自哈希函数映射后编码得到的哈希编码相似。CMSSH 通过标准的 AdaBoost 方法依次对每一位哈希编码学习得到对应的哈希函数,使得后一个哈希函数能够对前一个函数编码能力进行补充和加强。

作为最早的多视图哈希的算法,CMSSH 还存在一些不足,例如:在学习哈希函数的时候只考虑了不同类型的数据的相似度,而没有考虑同一类型数据的相似度;只适用于两种数据类型之间的跨越,不能扩展到通用的多种数据类型。

Kumar 等人提出了一种跨视图哈希索引方法(CVH)[94],并解决了 CMSSH 算法存在的两点不足。CVH 在进行哈希函数的学习时,提出了将同种模态间(intra-modality)和不同模态间(inter-modality)的两种相似度都纳入考虑中来进行哈希函数学习。哈希函数旨在最小化相近数据对应的哈希编码的距离同时最大化不相近数据的哈希编码的距离。Zhen 等人提出了一种基于概率图模型学习哈希函数的算法(MLBE)[95]。和以往利用谱图分割(spectral graph partition)来学习哈希函数的算法不同,MLBE 完全采用概率图模型来对学习得到哈希函数。对于来自不同模态的数据,相应的哈希函数将其由各自的原始高维空间中映射到一个二值隐空间(binary latent space)中,并保证模态内的相似性得以保持。同时,模态间相似性的约束施加在相应的隐空间之间,使得有语义关联的不同模态的数据具有相近的哈希编码。Zhen 等人又提出了一种协同正则的哈希函数学习方法(CRH)[96],旨在最小化多个模态的数据在哈希映射过程中的损失,从而学习得到每个模态数据最优的哈希函数。

在文献[97]中,Liu 等人利用统一超图(unified hypergraph)对 Flickr 网站中的图像及其上下文进行建模。在该模型中,图像和图像的上下文属性(如用户、标签、地理位置等)都被统一地看做超图中的节点,而异构数据间的高阶复杂关联则可以通过连接这些顶点的超边来自然表达。在此基础上,Liu 等人将谱哈希算法扩展到超图表达中,并借鉴自学习哈希(Self-taught Hashing)的思想将哈希算法应用到了新数据的计算。对来自不同模态的数据,分别利用各自模态学习到的哈希函数,可以映射为统一的哈希编码,从而进行高效的跨媒体检索。

跨媒体检索不仅局限于跨越不同模态的数据检索,将来自不同模态数据作为查询进行统一检索也属于跨媒体查询的范畴。Liu 等人提出了一种针对图像 + 关键词的混合查询的哈希索引[98]。在哈希函数学习的同时学习每个哈希函数和关键词的关联关系,在给出图像 + 关键词的查询条件时就可以先利用关键词找出最相关的哈希函数,并用这些哈希函数对图像进行编码,查找最相关的结果。

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 11, Nov. 2013

3.1.3 跨媒体排序

跨媒体排序的目的在于以更精细的粒度对索引返回集合里的跨媒体数据进行排序。跨媒体排序学习的目的在于用排序数据来学习两种不同模态之间的基于语义相似度的排序模型。

PAMIR(passive aggressive model for image retrieval)算法[99]首次提出用排序学习的方法对跨媒体图像进行排序。PAMIR 将跨媒体排序问题用一个类似于 RankSVM 的模型进行重新表达(因此它是个基于有序对的排序算法),并拓展了用于求解传统支持向量机的 Passive Aggressive 算法[100]以对海量的训练样本进行训练。同时,PAMIR 还可以结合核技巧(kernel trick)对图像的相似度进行不同的建模,从而选择在排序性能上表现最优的核函数。

在文献[101]中,Zhai 等人提出了 HSNN 方法用于计算不同模态媒体间的相似度以进行跨媒体排序。此异构的相似度是根据计算两个多媒体数据属于同一个语义类别的概率得来,而概率是根据分析异构每个多媒体数据的最近邻得来。在 HSNN 中,组合多个弱排序器并通过 AdaRank 来最终学习得到排序模型。

跨媒体排序中,文本和图像通常都用词袋模型(bag of words)或者视觉词袋模型(bag of visual words)来进行特征表达建模,因而往往被表达为高维特征向量。然而,高维向量空间表示引入了两个传统的问题,亦即"一词多义"和"一义多词"的问题。在传统文本检索中,为了捕获数据所包含的隐语义,将文本中的单词嵌入到一个低维隐空间是一种经典的解决方法,例如隐语义分析(LSA)[102]以及概率隐语义分析(Probabilistic LSA)[103]。类似的思想被引入到跨媒体排序中,如 SSI(supervised semantic indexing)方法[104]和 PSI(polynomial semantic indexing)方法[85]。SSI 方法定义了一类的低秩线性模型来捕获模态内以及模态间的"一词多义"和"一义多词"。与基于无监督学习的 LSI 不同的是,SSI 基于有监督的排序学习,使用了有序对作为训练样本。PSI 方法则扩展了 SSI,使之不仅能捕获线性关联,还能捕获更高阶的关联。

Lu 等人在文献[105]中提出以排序学习视角的跨媒体排序方法 LSCMR(latent semantic cross-modal ranking):LSCMR 通过两个不同的映射函数分别将两种不同模态的数据映射到同一个低维隐特征空间,使之可以在同一个特征空间里进行排序。和其他方法不同的是,LSCMR 基于最大化排序间隔的思想来学习得到这两个线性映射函数,因此能直接对最终的排序结果进行优化从而提高排序性能。

## 3.2 国内外研究进展比较

2012 年,国家科技部正式启动"面向公共安全的跨媒体计算理论与方法"的国家重点基础研究发展计划(973)项目(项目编号:2012CB316400),这一项目拟对跨媒体数据统一表示与建模、跨媒体数据关联推理与深度挖掘和跨媒体数据综合搜索和内容合成等 3 个方面进行重点研究,面向社会公共安全这一重大需求,对跨媒体数据所隐含的社会属性和特定事件的动态演变进行分析,挖掘和预测特定社会事件,进而以跨媒体形式对其起源、现状及发展进行全过程呈现。

对不同类型媒体数据的综合利用及搜索也成为近年来国际上的一个热点趋势。2010 年 1 月,自然(Nature)杂志邀请 20 位搜索、医药、能源、天文学、化学、人类健康、激光、石油、生态学的专家和政策制定者对下一个 10 年科技发展进行了展望,形成的报告《2020 Vision》发表于自然杂志 2010 年第 1 期。Google 的研究主管 Norvig 博士应邀写了对未来多媒体计算与搜索技术的预测:文本、图像、视频将更加紧密混合在一起(跨媒体),亲朋好友之间交往的社会化属性等信息将被记录和分析(社会网络),传感器网络在定位、医疗等方面应用更加普遍(物联网)。搜索请求不再是文本,而是语音和对大脑信号的理解(语音识别与人脑感知);搜索结果不再是罗列网页,而是以图或表来表示的更为形象的综合性知识(数据挖掘)。国际期刊《Journal of Multimedia》于 2012 年 2 月推出了名为"Connected Multimedia"的特刊,讨论了多模态数据主题建模、基于多模态数据分类的路线推荐等问题,《Journal of Zhejiang University—Science C》也于 2012 年 12 月推出了名为"Societally connected multimedia across cultures"的特刊,对类似问题进行了深入讨论。同时,国外著名研究机构 INRIA 成立了关于多模态数据的内容检索项目"I-Search",对来自各种媒体多模态的数据(文本、图像、视频、音频、3D 模型等)进行统一检索。

总体而言,国内外对跨媒体检索理论并方法均抱以很大热情并投入很大精力进行研究,国内外在这一方向的研究处于同步阶段。但是,在跨媒体度量学习、索引和排序等方面尚存在巨大挑战。在跨

媒体度量方面,目前的手段主要有典型相关性分析、跨模态字典学习和概率主题建模,它们分别从相关度学习、共享结构挖掘和稀疏重建等角度来解决跨媒体度量学习,因此,如何将三者结合起来,构建基于概率隐模型的跨媒体度量学习方法,是一个值得深入研究的问题;在跨媒体索引方面,哈希的方法是最近的热门研究方向,如何将来自不同模态的数据映射为统一的简短的哈希编码,从而实现高效的跨媒体检索是跨媒体哈希索引的目标。因此,如何对数据进行哈希编码的同时,保持模态内的相似性和模态间的相关性是至关重要的因素,值得深入探索;在跨媒体排序方面,目前的工作主要是基于相关性的排序,即输入一个模态的查询数据,返回和它相关的另一个模态的有序的反馈结果。此外,在此基础上还有一些值得进一步深化的研究问题,例如:支持更为复杂的查询模式如复合模态的查询输入,或者返回多个模态的相关结果,并进行统一排序等。

### 3.3 发展趋势与展望

在谈及跨媒体这一概念时,我们需要分清其与多媒体这一概念存在以下两个方面的不同:

1)跨媒体数据强调不同类型数据来源于不同渠道(sources 或 domains),在分析这些数据时要关注 3 个方面的"跨越(cross)",即跨越模态(cross-modality)、跨越来源(cross-source)和跨越网络空间和实体世界(cross cyberspace to reality)。跨越模态指在检索过程中可以利用不同类型的异构底层特征(如文本特征和视觉特征等)来提升检索性能;跨越来源指不同来源数据(如从 Flickr 和 Twitter 获取的数据)从不同侧面来表达语义;跨越空间指跨媒体数据通过人类交互行为来反映网络空间和实体世界的相互耦合、相互补充。

2)与第一点相一致的是在跨媒体分析过程中我们需要充分利用从不同来源所获取的不同类型数据来挖掘其蕴含的知识,如文本和图像数据中所内嵌的隐性关联、不同类型数据合成在一起作为摘要来表达多源头数据的内容侧面(aspect)等等。因此,如何去发现和挖掘从不同源头所获取的不同类型数据之间的关联性是跨媒体研究所面临的巨大挑战,正如加州大学伯克利分校统计系前任系主任 Speed 教授于 2011 年 12 月在《Science》发表的题为"A correlation for the 21st century"的论文所提出的"21 世纪是关联性学习的时代"这一观点,在跨媒体检索中,从庞大数据集中发现数据之间所潜在的重要关系变得十分重要。

## 4 结 论

对媒体信息的索引和检索的终极目标是实现类似文本搜索的搜索引擎,通过文本可以找到任何想要的多媒体内容。在视频智能管理和智能视觉监控需求的推动下,事件识别与标注分析技术蓬勃发展,虽然有很多的系统实现了对简单事件的分析与理解,但是视频中复杂场景下复杂的事件分析一直没有得到很好地解决。随着数据量的增加,现有的高维索引结构仍然无法彻底克服"维度灾难",无法达到文本倒排索引的效果,更无法满足搜索引擎的实际需求。跨媒体搜索为跨越"语义鸿沟"提供了很好的解决思路并已经取得很好的效果,但远远无法满足基于内容搜索的需要。随着大数据时代的到来,多媒体信息的索引与检索的需求将日益迫切,面对众多挑战的同时,该研究领域将迎来前所未有的重大机遇,将会有越来越多的研究者关注该领域,也必将产生越来越多可以实用的研究成果。

## 志 谢

本文视频事件检测与标注部分由中国科学院自动化所模式识别国家重点实验室的徐常胜研究员和张天柱博士撰写,高维索引结构部分由华中科技大学计算机学院的于俊清教授和艾列富博士撰写,跨媒体检索理论与技术部分由浙江大学计算机学院的庄越挺教授和吴飞教授撰写,于俊清负责统稿。本文在选题和内容上得到了计算机学会多媒体专业委员会的大力支持和悉心指导。

### 参考文献(References)

[1] Aggarwal J, Cai Q. Human motion analysis: a review [C]//Proceedings of IEEE Nonrigid and Articulated Motion Workshop. New York: IEEE Press, 1997: 90-102. [DOI: 10.1109/NAMW.1997.609859]

[2] Gavrila D. The visual analysis of human movement: a survey [J]. Computer Vision and Image Understanding, 1999, 73(1): 82-98. [DOI: 10.1006/cviu.1998.0716]

[3] Wang J J, Singh S. Video analysis of human dynamics-a survey [J]. Real-Time Imaging, 2003, 9(5): 321-346. [DOI: 10.1016/j.rti.2003.08.001]

[4] Wang L, Hu W, Tan T. Recent developments in human motion

analysis [J]. Pattern Recognition, 2003, 36(3): 585-601. [DOI: 10.1016/S0031-3203(02)00100-0]

[5] Hu W, Tan T, Wang L, et al. A survey on visual surveillance of object motion and behaviors [J]. IEEE Systems, Man, and Cybernetics Society, 2004, 349(3): 334-352. [DOI: 10.1109/TSMCC.2004.829274]

[6] Ronald P. Vision-based human motion analysis: an overview [J]. Computer Vision and Image Understanding, 2007, 108(1):4-18. [DOI: 10.1016/j.cviu.2006.10.016]

[7] Turaga P, Chellappa R, Subrahmanian V S, et al. Machine recognition of human activities: A survey [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(11): 1473-1488. [DOI: 10.1109/TCSVT.2008.2005594]

[8] Poppe R. A survey on vision-based human action recognition [J]. Image and Vision Computing, 2010, 28(6): 976-990. [DOI: 10.1016/j.imavis.2009.11.014]

[9] Bobick A, Davis J. The recognition of human movement using temporal templates [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2001, 23(3): 257-267. [DOI: 10.1109/34.910878]

[10] Blank M, Gorelick L, Shechtman E, et al. Actions as space-time shapes [C]//Proceedings of ICCV. New York: IEEE Press, 2005: 1395-1402. [DOI: 10.1109/ICCV.2005.28]

[11] Ben A, Wang Z, Pandit P, et al. Human activity recognition using multidimensional indexing [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002, 24(8): 1091-1104. [DOI: 10.1109/TPAMI.2002.1023805]

[12] Wada T, Matsuyama T. Multi object behavior recognition by event driven selective attention method [J]. IEEE TPAMI, 2000, 22(8): 873-887. [DOI: 10.1109/34.868687]

[13] Niebles C, Wang H, Li F. Unsupervised learning of human action categories using spatial-temporal words [J]. International Journal of Computer Vision, 2008, 79(3): 299-318. [DOI: 10.1007/s11263-007-0122-4]

[14] Wang Y, Jiang H, Drew M S, et al. Unsupervised discovery of action classes [C]//Proceedings of CVPR. New York: IEEE Press, 2006: 1654-1661. [DOI: 10.1109/CVPR.2006.321]

[15] Dedeoglu Y, Toreyin B, Gudukbay U, et al. Silhouette-based method for object classification and human action recognition in video [C]//Proceedings of HCI. Berlin, Heidelberg: Springer, 2006: 64-77. [DOI: 10.1007/11754336_7]

[16] Daniel W, Edmond B. Action recognition using exemplar-based embedding [C]//Proceedings of CVPR. New York: IEEE Press, 2008: 1-7. [DOI: 10.1109/CVPR.2008.4587731]

[17] Du T, Alexander S. Human activity recognition with metric learning [C]//Proceedings of ECCV. Berlin, Heidelberg: Springer, 2008: 548-561. [DOI: 10.1007/978-3-540-88682-2_42]

[18] VenkateshBabu R, Ramakrishnan K R. Recognition of human actions using motion history information extracted from the compressed video [J]. Image and Vision Computing, 2004, 22(8): 597-607. [DOI: 10.1016/j.imavis.2003.11.004]

[19] Carlsson S, Sullivan J. Action recognition by shape matching to key frames [C]//Proceedings of Workshop on Models Versus Exemplars in Computer Vision. New York: IEEE Press, 2001: 1-8.

[20] Laptev I, Lindeberg T. Velocity adaptation of space-time interest points [C]//Proceedings of ICPR. New York: IEEE Press, 2004: 1-4. [DOI: 10.1109/ICPR.2004.1334003]

[21] Brand M, Kettnaker V. Discovery and segmentation of activities in video [J]. IEEE TPAMI, 2000, 22(8): 844-851. [DOI: 10.1109/34.868685]

[22] Hongeng S, Nevatia R. Large-scale event detection using semi-hidden Markov models [C]//Proceedings of ICCV. New York: IEEE Press, 2003: 1455-1462. [DOI: 10.1109/ICCV.2003.1238661]

[23] Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequential images using hidden Markov model [C]//Proceedings of CVPR. New York: IEEE Press, 1992: 379-385. [DOI: 10.1109/CVPR.1992.223161]

[24] Starner T, Pentland A. Visual recognition of american sign language using hidden Markov models [C]// Proceeding of International Workshop on Automatic Face and Gesture Recognition. Zurich: University of Zurich, 1995: 189-194.

[25] Weinland D, Boyer E, Ronfard R. Action recognition from arbitrary views using 3D exemplars [C]//Proceedings of ICCV. New York: IEEE Press, 2007: 1-7. [DOI: 10.1109/ICCV.2007.4408849]

[26] Luo Y, Wu T, Hwang J. Object based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks [J]. Computer Vision and Image Understanding, 2003, 92(2-3): 196-216. [DOI: 10.1016/j.cviu.2003.08.001]

[27] Ghahramani Z, Jordan M. Factorial hidden Markov model [J]. Machine Learning, 1997, 29(2): 245-273. [DOI: 10.1023/A:1007425814087]

[28] Bui H, Venkatesh S, West G. Policy recognition in the abstract hidden Markov model [J]. Journal of Artificial Intelligence Research, 2002, 17: 451-499. [DOI: 10.1613/jair.839]

[29] Michael S R, Aggarwal J K. Recognition of composite human activities through context-free grammar based representation [C]// Proceedings of CVPR. New York: IEEE Press, 2006: 1709-1718. [DOI: 10.1109/CVPR.2006.242]

[30] Hu W, Xie T D. A hierarchical self-organizing approach for learning the patterns of motion trajectories [J]. IEEE Transactions on Neural Networks, 2004, 15(1): 135-144. [DOI: 10.1109/TNN.2003.820668]

[31] Vail D L, Manuela M V, Lafferty J D. Conditional random fields for activity recognition [C]//Proceedings of AAMAS. New York: ACM Press, 2007: 1-8. [DOI: 10.1145/1329125.1329409]

[32] Sminchisescu C, Kanaujia A, Li Z, et al. Conditional models for

contextual human motion recognition [C]//Proceedings of ICCV. New York: IEEE Press, 2005: 1808-1815. [DOI: 10.1109/ICCV.2005.59]

[33] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies [C]//Proceedings of CVPR. New York: IEEE Press, 2008: 1-8. [DOI: 10.1109/CVPR.2008.4587756]

[34] Liu J, Luo J, Shah M. Recognizing realistic actions from videos "in the wild" [C]//Proceedings of CVPR. New York: IEEE Press, 2009: 1996-2003. [DOI: 10.1109/CVPR.2009.5206744]

[35] Zhu G, Yang M, Yu K, et al. Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor [C]//Proceedings of ACM Multimedia. New York: ACM Press, 2009: 165-174. [DOI: 10.1145/1631272.1631297]

[36] Xiang T, Gong S. Beyond tracking: modelling activity and understanding behavior [J]. IJCV, 2006, 67(1): 21-51. [DOI: 10.1007/s11263-006-4329-6]

[37] Manor Z, Irani M. Event-based analysis of video [C]//Proceedings of CVPR. New York: IEEE Press, 2001: 123-130. [DOI: 10.1109/CVPR.2001.990935]

[38] Zhang D, Perez D G, Bengio S, et al. Semi-supervised adapted hmms for unusual event detection [C]//Proceedings of CVPR. New York: IEEE Press, 2005: 611-618. [DOI: 10.1109/CVPR.2005.316]

[39] Chanop S A, Richard H. Optimised KD-trees for fast image descriptor matching [C]//Proceedings of CVPR. New York: IEEE Press, 2008: 1-8. [DOI: 10.1109/CVPR.2008.4587638]

[40] Zhuang Y, Zhuang Y, Li Q, et al. Indexing high-dimensional data in dual distance spaces: a symmetrical encoding approach [C]//Proceedings of the Conference on Extending Database Technology. New York: ACM Press, 2008: 241-251. [DOI: 10.1145/1353343.1353375]

[41] He Y, Yu J. MFI-Tree: an effective multi-feature index structure for weighted query application [J]. Computer Science and Information System, 2010, 7(1): 139-151. [DOI: 10.2298/CSIS1001139H]

[42] Mayur D, Nicole I, Piotr I, et al. Locality-sensitive hashing scheme based on p-stable distributions [C]//The Annual Symposium on Computational Geometry. New York: ACM Press, 2004: 253-262. [DOI: 10.1145/997817.997857]

[43] Alexandr A, Piotr I. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions [J]. Communications of the ACM, 2008, 51(1): 117-122. [DOI: 10.1145/1327452.1327494]

[44] Kuo Y, Chen K, Chiang C, et al. Query expansion for hash-based image object retrieval [C]//Proceedings of ACM Multimedia. New York: ACM Press, 2009: 65-74. [DOI: 10.1145/1631272.1631284]

[45] Rina P. Entropy based nearest neighbor search in high dimensions [C]//ACM-SIAM Symposium on Discrete Algorithm. New York: ACM Press, 2006: 1186-1195. [DOI: 10.1145/1109557.1109688]

[46] Lv Q, Josephson W, Wang Z, et al. Multi-probe LSH: efficient indexing for high-dimensional similarity search [C]//Proceedings of VLDB. New York: ACM Press, 2007: 950-961. [DOI: 978-1-59593-649-3]

[47] Herve J, Laurent A, Cordelia S, et al. Query-adaptative locality sensitive hashing [C]//Proceedings of ICASSP. New York: IEEE Press, 2008: 825-828. [DOI: 10.1109/ICASSP.2008.4517737]

[48] Loic P, Herve J, Laurent A. Locality sensitive hashing: a comparison of hash function types and querying mechanisms [J]. Pattern Recognition Letters, 2010, 31(11): 1348-1358. [DOI: 10.1016/j.patrec.2010.04.00]

[49] Antonio T, Rob F, Yair W. Small codes and large image databases for recognition [C]//Proceedings of CVPR. New York: IEEE Press, 2008: 1-8. [DOI: 10.1109/CVPR.2008.4587633]

[50] Yair W, Antonio T, Rob F. Spectral hashing [C]//Proceedings of NIPS. British, Columbia: NIPS Foundation, 2008: 1753-1760.

[51] He J, Regunathan R, Chang S, et al. Compact hashing with joint optimization of search accuracy and time [C]//Proceedings of CVPR. New York: IEEE Press, 2011: 753-760. [DOI: 10.1109/CVPR.2011.5995518]

[52] Alexis J, Olivier B. Random maximum margin hashing [C]//Proceedings of CVPR. New York: IEEE Press, 2011: 873-880. [DOI: 10.1109/CVPR.2011.5995709]

[53] Heo J, Lee Y, He J, et al. Spherical hashing [C]//Proceedings of CVPR. New York: IEEE Press, 2012: 2957-2964. [DOI: 10.1109/CVPR.2012.6248024]

[54] Christoph S, Alexander M B. LDAHash: improved matching with smaller descriptors [J]. IEEE TPAMI, 2012, 34(1): 66-78. [DOI: 10.1109/TPAMI.2011.103]

[55] Mohammad N, David J F. Minimal loss hashing for compact binary codes [C]//Proceedings of Conference on Machine Learning. New York: Association for Computing Machinery, 2011: 1-8.

[56] Wang J, Kumar S, Chang S. Semi-supervised hashing for large scale search [J]. IEEE TPAMI, 2012, 34(12): 2393-2406. [DOI: 10.1109/TPAMI.2012.48]

[57] Liu W, Wang J, Ji R, et al. Supervised hashing with kernels [C]//Proceedings of CVPR. New York: IEEE Press, 2012: 2074-2081. [DOI: 10.1109/CVPR.2012.6247912]

[58] Shao J, Wu F, Ouyyang C, et al. Sparse spectral hashing [J]. Pattern Recognition Letters, 2012, 33(1): 271-277. [DOI: 10.1016/j.patrec.2011.10.018]

[59] Zhuang Y, Liu Y, Wu F, et al. Hypergraph spectral hashing for similarity search of social image [C]//Proceedings of ACM Multimedia. New York: ACM Press, 2011: 1457-1460. [DOI:

10. 1145/2072298. 2072039]

[60] Zhou W, Lu Y, Li H, et al. Scalar quantization for large scale image search [C]//Proceedings of the ACM Multimedia. New York: ACM Press, 2012: 169-178. [DOI: 10. 1145/2393347. 2393377]

[61] Sivic J, Zesserman A. Video google: a text retrieval approach to object matching in video [C]//Proceedings of ICCV. New York: ACM Press, 2003: 1470-1477. [DOI: 10. 1109/ICCV. 2003. 1238663]

[62] David N, Henrik S. Scalable recognition with a vocabulary tree [C]//Proceedings of CVPR. New York: IEEE Press, 2006: 2161-2168. [DOI: 10. 1109/CVPR. 2006. 264]

[63] James P, Ondrej C, Michael I, et al. Object retrieval with large vocabularies and fast spatial matching [C]//Proceedings of CVPR. New York: IEEE Press, 2007: 1-8. [DOI: 10. 1109/CVPR. 2007. 383172]

[64] Mu Y, Sun J, Han T X, et al. Randomized locality sensitive vocabularies for bag-of-features model [C]//Proceedings of ECCV. Berlin, Heidelberg: Springer-Verlag, 2010: 748-761. [DOI: 10. 1007/978-3-642-15558-1_54]

[65] Kuo Y, Lin H, Cheng W, et al. Unsupervised auxiliary visual words discovery for large-scale image object retrieval [C]//Proceedings of CVPR. New York: IEEE Press, 2011: 905-912. [DOI: 10. 1109/CVPR. 2011. 5995639]

[66] Jan C G, Cor J V, Arnold W M S. Visual word ambiguity [J]. IEEE TPAMI, 2010, 32(7): 1271-1283. [DOI: 10. 1109/TPAMI. 2009. 132]

[67] James P, Ondrej C, Michael I, et al. Lost in quantization: improving particular object retrieval in large scale image databases [C]//Proceedings of CVPR. New York: IEEE Press, 2008: 1-8. [DOI: 10. 1109/CVPR. 2008. 4587635]

[68] Ai L, Yu J, Guan T. Spherical soft assignment: improving image representation in content-based image retrieval [C]// Proceeding of PCM. Berlin, Heidelberg: Springer-Verlag, 2012, 801-810. [DOI: 10. 1007/978-3-642-34778-8_75]

[69] Jegou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C]//Proceedings of CVPR. New York: IEEE Press, 2010: 3304-3311. [DOI: 10. 1109/CVPR. 2010. 5540039]

[70] Chen D, Tsai S, Chandrasekhar V, et al. Residual enhanced visual vectors for on-device image matching [C]//Proceedings of ASILOMAR. New York: IEEE Press, 2011: 850-854. [DOI: 10. 1109/ACSSC. 2011. 6190128]

[71] Jegou H, Douze M, Schmid C. Improving bag-of-feature for large scale image search [J]. IJCV, 2010, 87(3): 316-336. [DOI: 10. 1007/s11263-009-0285-2]

[72] Jegou H, Douze M, Schmid C. Packing bag-of-features [C]// Proceedings of ICCV. New York: IEEE Press, 2009: 2357-2364. [DOI: 10. 1109/ICCV. 2009. 5459419]

[73] Jain M, Jegou H, Gros P. Asymmetric hamming embedding: taking the best of our bits for large scale image search [C]//Proceedings of ACM Multimedia. New York: ACM Press, 2011: 1441-1444. [DOI: 10. 1145/2072298. 2072035]

[74] Jegou H, Douze M, Schmid C. Product quantization for nearest neighbor search [J]. IEEE TPAMI, 2011, 33(1): 117-128. [DOI: 10. 1109/TPAMI. 2010. 57]

[75] Brandt J. Transform coding for fast approximate nearest neighbor search in high dimensions [C]//Proceedings of CVPR. New York: IEEE Press, 2010: 1815-1822. [DOI: 10. 1109/CVPR. 2010. 5539852]

[76] Chen Y, Guan T, Wang C. Approximate nearest neighbor search by residual vector quantization [J]. Sensors, 2010, 10(12): 11259-11273. [DOI: 10. 3390/s101211259]

[77] Babenko A, Lempitsky V. The inverted multi-index [C]//Proceedings of CVPR. New York: IEEE Press, 2012: 3069-3076. [DOI: 10. 1109/CVPR. 2012. 6248038]

[78] Hotelling H. Relations between two sets of variates [J]. Biometrika, 1936, 28(3/4): 321-377.

[79] Hardoon D R, Taylor J S. Sparse canonical correlation analysis [J]. Machine Learning, 2011, 83(3): 331-353. [DOI: 10. 1007/s10994-010-5222-7]

[80] Chen X, Liu H, Carbonell J G. Structured sparse canonical correlation analysis [C]//Proceedings of the 15th International Conference on Artificial Intelligence and Statistics. La Palma: AI & Statistics committee, 2012: 199-207.

[81] Rupnik J, Taylor J S. Multi-view canonical correlation analysis [C]//Proceedings of the Conference on Data Mining and Data Warehouses. Ljubljana: SiKDD Committee, 2010: 1-4.

[82] Sharma A, Kumar A, Daume H, et al. Generalized multiview analysis: a discriminative latent space [C]//Proceedings of CVPR. New York: IEEE Press, 2012: 2160-2167. [DOI: 10. 1109/CVPR. 2012. 6247923]

[83] Zhuang L, Zhuang Y, Wu J, et al. Image retrieval approach based on sparse canonical correlation analysis [J]. Journal of Software, 2012, 23(5): 1295-1304. [DOI: 10. 3724/SP. J. 1001. 2012. 04032]

[84] Jia Y, Salzmann M, Darrell T. Factorized latent spaces with structured sparsity [C]//Proceedings of NIPS. Vancouver: NIPS Foundation, 2010: 982-990.

[85] Bai B, Weston J, Grangier D, et al. Polynomial semantic indexing [C]//Proceedings of NIPS. New York: Curran Associates, 2009: 1-9.

[86] Putthividhy D, Attias H T, Nagarajan S S. Topic regression multi-modal latent dirichlet allocation for image annotation [C]//Proceedings of CVPR. New York: IEEE Press, 2010: 3408-3415. [DOI: 10. 1109/CVPR. 2010. 5540000]

[87] Jia Y, Salzmann M, Darrell T. Learning cross-modality similarity for multinomial data [C]//Proceedings of ICCV. New York: IEEE Press, 2011: 2407-2414. [DOI: 10. 1109/ICCV. 2011. 6126524]

[88] Virtanen S, Jia Y, Klami A, et al. Factorized multi-modal topic model, UAI-P-2012-PG-843-851 [R]. Ithaca: Cornell Universi-

ty Library, 2012 [arXiv:1210.4920]

[89] Zhuang Y, Wang Y, Wu F, et al. Supervised coupled dictionary learning with group structures for multi-modal retrieval [C]// Proceedings of the 27th Conference on Artificial Intelligence. Washington, USA: AAAI organization, 2013: 1070-1076.

[90] Yu Z, Tang S, Zhang Y, et al. Image ranking via attribute boosted hypergraph [C]//Proceedings of PCM. Berlin, Heidelberg: Springer-Verlag, 2012: 779-789. [DOI: 10.1007/978-3-642-34778-8_73]

[91] Hong C, Zhu J. Hypergraph-based multi-example ranking with sparse representation for transductive learning image retrieval [J]. Neurocomputing, 2013, 101(4): 94-103. [DOI: 10.1016/j.neucom.2012.09.00]

[92] Hong C, Jun Y, Jonathan L, et al. Multi-view hypergraph learning by patch alignment framework [J]. Neurocomputing, 2013, 118(22): 79-86. [DOI: 10.1016/j.neucom.2013.02.017]

[93] Bronstein M M, Bronstein A M, Michel F, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing [C]//Proceedings of CVPR. New York: IEEE Press, 2010: 3594-3601. [DOI: 10.1109/CVPR.2010.5539928]

[94] Kumar S, Udupa R. Learning hash functions for cross-view similarity search [C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. New York: ACM Press, 2011: 1360-1365. [DOI: 10.5591/978-1-57735-516-8/IJCAI11-230]

[95] Zhen Y, Yeung D Y. A probabilistic model for multimodal hash function learning [C]//Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 940-948. [DOI: 10.1145/2339530.2339678]

[96] Zhen Y, Yeung D Y. Co-regularized hashing for multimodal data [C]//Proceedings of NIPS. Lake Tahoe: NIPS Foundation, 2012: 1385-1393.

[97] Liu Y, Shao J, Wu F, et al. Hypergraph spectral hashing for image retrieval with heterogeneous social contexts [J]. Neurocomputing, 2013, 119(7): 49-58. [DOI: 10.1016/j.neucom.2012.02.051]

[98] Liu X, Mu Y, Lang B, et al. Compact hashing for mixed image-keyword query over multi-label images [C]//Proceedings of ICMR. New York: ACM Press, 2012: 18-27. [DOI: 10.1145/2324796.2324819]

[99] Grangier D, Bengio S. A discriminative kernel-based model to rank images from text queries [J]. IEEE TPAMI, 2008, 30(8): 1371-1384. [DOI: 10.1109/TPAMI.2007.70791]

[100] Crammer K, Dekel O, Keshet J, et al. Online passive-aggressive algorithms [J]. The Journal of Machine Learning Research, 2006, 7(12): 551-585.

[101] Zhai X, Peng Y, Xiao J. Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval [C]//Proceedings of Advances in Multimedia Modeling. Berlin, Heidelberg: Springer-Verlag, 2012: 312-322. [DOI: 10.1007/978-3-642-27355-1_30]

[102] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[103] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning, 2001, 42(1-2): 177-196. [DOI: 10.1023/A:1007617005950]

[104] Bai B, Weston J, Grangier D. Learning to rank with(a lot of) word features [J]. Information Retrieval, 2009, 13(3): 291-314. [DOI: 10.1007/s10791-009-9117-9]

[105] Lu X, Wu F, Tang S, et al. A low rank structural large margin method for cross-modal ranking [C]//Proceedings of the 36th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval. New York: ACM Press, 2013: 433-442. [DOI: 10.1145/2484028.2484039]