国家重点基础研究发展计划（973 计划）项目

“面向公共安全的跨媒体计算理论与方法”简报

（2013 年第 2 期）

项目管理办公室编制　　2013 年 8 月 1 日

“面向公共安全的跨媒体计算理论与方法”973 项目北京研讨会举行

2013 年 4 月 2 日，国家重点基础研究发展计划（973 计划）项目“面向公共安全的跨媒体计算理论与方法”（项目编号：2012CB316400）在北京召开了项目研讨会。高文院士、袁保宗教授、施鹏飞教授、戴国忠教授、钱德沛教授、李波教授、胡事民教授等专家出席了本次报告会。

中科院计算所陈熙霖副所长致报告会开幕辞，陈熙霖研究员认为 973 项目提供了一个极好的交流学习机会，促进了各个学校、研究所和项目团队间彼此的相互了解，他预祝项目研讨会取得圆满成功。

项目首席科学家庄越挺教授代表整个项目组对 2012 年项目整体进展情况向专家进行了简略汇报，包括项目研究内容与关键技术、课题设置与分工、项目拟采用的技术路线以及需要解决的三个关键科学问题。庄老师随后进一步汇报了自 2012 年立项以来各课题研究的基本情况，包括发表论文以及在团队建设、项目管理、国际合作等方面情况。其中特别提到 2013 年多媒体领域顶级学术会议 ACM Multimedia 2013 在其 multimedia analysis track 中列出了 cross-media analysis 方向，这显示“跨媒体”研究已经逐渐被国际学术界所认同。庄老师的汇报还提及了项目组课题间交流研讨情况、项目简报、示范平台搭建等其他相关方面的情况。

朱振峰老师围绕课题一“跨媒体数据统一表示和建模”的进展进行了汇报。汇报内容主要围绕基元提取、一致性表示、关联分析与增量索引这四个方面。针对跨媒体基元提取与描述，课题一提出了结合词典学习去提取具有不变性的跨媒体基元元素的方法；针对跨媒体语义一致性描述，课题一提出了广义 CCA （GCCA）的方法来挖掘跨媒体数据共享特征；针对跨媒体关联建模，进行了跨模态数据关联、跨“维度”的媒体数据关联建模、基于 Factor Graph Model 的音视情感建模、基于 DNN 的音视频联合建模等研究；针对增量整合和高效索引，研究了基元空间分割的词典构造及映射方法。

薛建儒教授代表课题二“跨媒体属性感知模型与行为计算”进行了汇报。汇报内容主要围绕跨媒体数据的有效性辨识、跨媒体属性感知、视觉显著性计算模型、社会关注度模型等方面。在跨媒体数据的有效性辨识方面，提出了结合跨媒体数据的自然属性和社会属性来实现

跨媒体数据有效性检测算法，对存在矛盾、不真实的跨媒体数据进行甄别；在跨媒体属性感知方面，通过借鉴视觉认知机理，构造跨媒体数据的注意力模型，实现跨媒体数据的属性感知，并基于属性建立不同类型数据之间的关联，为社会热点事件发现与推演提供支撑；在社会关注度模型方面，通过挖掘跨媒体数据自然属性与社会属性的分布特征，建立社会关注度模型的方法。

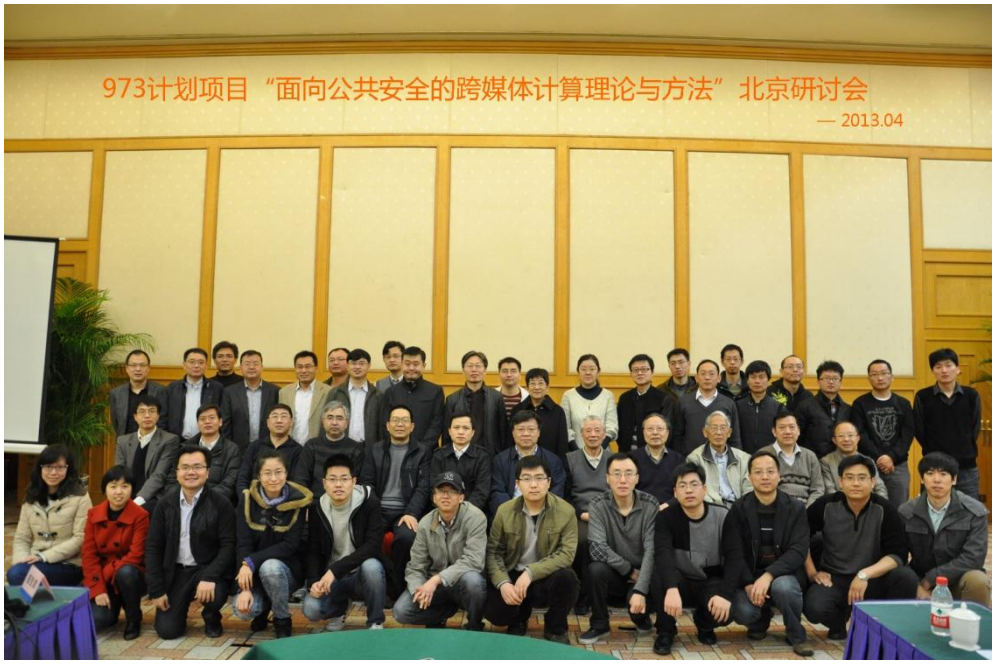
课题三负责人计算所黄庆明教授汇报课题三“跨媒体语义学习与内容理解”的研究进展汇报，内容主要围绕跨媒体语义单元学习和标注、话题及事件的结构模式表示与检测这两个方面。针对跨媒体语义单元学习和标注，研究了多特征融合及多核学习、视觉语义关联学习、层次化字典和判别模型学习、基于空域信息的稀疏编码等方法；针对话题及事件的结构模式表示与检测，研究了利用互增强的视频紧凑表示、结合语义共生性及时域突发模型的话题检测、基于协同聚类和多约束的话题检测与关联、基于多源信息融合与图聚类的话题检测等方法。

何晓飞教授随后代表课题四“跨媒体语义学习与内容理解”进行了工作汇报，其汇报内容主要围绕海量跨媒体数据高效处理与跨媒体数据关联挖掘这两个方面。针对海量跨媒体数据高效处理，课题四重点研究跨媒体数据几何拓扑性质的分析，超大规模数据的快速代表性采样，大规模矩阵的快速高效分解等问题。其近期的成果表现在海量跨媒体数据流形学习理论与海量跨媒体数据的主题建模；在跨媒体数据关联挖掘方面，课题四重点研究了基于平行向量场的降维理论，提出了基于平行向量场的排序、基于平行向量场的跨媒体多任务学习等方法；在跨媒体数据的主题建模方面，研究了基于带限制的非负矩阵分解的图像表示理论与基于 A-最优非负投影的图像表示理论。

吴飞教授代表课题五“跨媒体搜索与内容整合”进行了汇报，汇报内容主要包括跨媒体索引、跨媒体度量学习、跨媒体排序、跨域摘要生成等方面。针对跨媒体哈希索引，主要研究同时生成不同类型媒体数据紧凑的汉明哈希编码方法；针对跨媒体度量学习，主要研究学习不同类型媒体数据之间相似性(如文本和图像之间的相似度)，提出了基于组结构的监督耦合字典学习与基于隐条件随机场的跨媒体度量学习；针对跨媒体排序，主要研究给定一种类型的查询数据，生成另外一种类型数据的排序列表（如用文本作为检索，得到排序的图像数据），提出了基于低秩结构最大间隔学习的跨媒体排序与双向学习的跨媒体排序方法；针对跨域摘要生成，研究了从不同来源、不同类型数据出发，生成描述同一主题的结构性摘要。

于慧敏教授代表课题六“面向公共安全的跨媒体呈现与验证和示范平台”进行工作进展汇报，其汇报的内容包括跨媒体的非完美标注方法、主题的演变/演化与因果推断理论模型和解决方案、事件因果关系的辩护与预测、跨媒体计算验证平台等四个方面。针对跨媒体的非完美标注方法，课题六重点研究了跨媒体中的多视图非完美标注学习方法；针对跨媒体话题网络学习与推理模型，主要研究了跨媒体话题网络结构学习与参数学习、话题推理、跨媒体话题提取模型、主题演化与跟踪模型、基于抽象论辩理论的事件因果关系的辩护与预测等方法；针对跨媒体计算验证平台，设计了平台架构，建立基于对象、事件、主题的跨媒体语义表达模型，集成跨媒体挖掘算法，支持跨媒体检索。在验证平台方面，课题六汇聚了主流网站中网页、微博、图像、视频等数据，构建了跨媒体数据集；按照“对象、事件、主题”的跨媒体语义表达模型，完成数据的组织与存储；完成了验证平台的界面设计。

与会人员展开了热烈讨论，项目咨询组责任专家和项目组专家提出了诸多建议。



“面向公共安全的跨媒体计算理论与方法”973 项目北京研讨会代表合影（2013 年 4 月 2 日）

首席科学家庄越挺教授在最后做了总结：他认为专家分析的非常到位，非常宝贵，我们开这次会的目的就是得到专家的宝贵建议，以便我们下阶段可以做得更好。各课题在研究中要重点突出 “跨”字，更加主动本 973 项目要突破的关键理论问题上靠，要突出重点，工作要有所区分，要落实专家意见。

项目受邀为中国计算机学会撰写年度发展报告跨媒体检索理论与技术相关内容

注：中国计算机学会多媒体技术专业委员会在撰写 2012 年度进展报告时，邀请项目组庄越挺教授和吴飞教授撰写跨媒体检索理论与技术相关内容。现将该技术报告中涉及跨媒体度量、跨媒体索引和跨媒体排序等方面的内容概括总结如下，详细内容可参见技术报告。

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。典型相关性分析（Canonical Correlation Analysis, CCA）是一种最早被用来实现不同类型数据之间检索的方法，其后出现了核化 CCA、稀疏 CCA、结构稀疏 CCA 以及泛化多视图分析（Generalized Multiview Analysis，GMA）等方法。为了充分利用数据中的“隐藏结构（如主题）”， LDA(Latent Dirichlet Allocation)也被用来实现对多模态（多类型）数据进行度量学习，如 Correspondence LDA (Corr-LDA)。随着字典学习这一方法的兴起，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，如多模态字典学习等。

2） 跨媒体索引

现今的多媒体数据哈希索引的研究方向大致可以划分为如下三类，即：

单一类型特征哈希索引：指以单一类型的高维特征为输入的一类哈希算法的总称；多视图哈希索引：对从数据中提取的不同类型特征和属性等进行索引的方法可归纳为多视图索引；跨媒体哈希索引：对包含异构类型数据的信息资源进行哈希索引可称为跨媒体哈希索引。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。目前代表性的跨媒体索引方法有跨媒体哈希索引（CMSSH）、跨视图哈希索引、基

于概率图模型学习哈希函数（Multimodal latent binary embedding，MLBE）、协同正则的哈希函数学习（Co- Regularized Hashing，CRH）等方法。

3）跨媒体排序

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。跨媒体排序方法的假设是，与查询词越相关的图片会被用户点击得越多。因此，它们都在寻求一个解放方案：如何更好地建立跨媒体数据之间的关联，使得对于同一查询词，被用户点击越多的图像（也就是越相关的图片）被排序得越靠前，且这种关联关系可以推广到未知的查询词和未知的检索文档上。需要强调的是跨媒体排序学习和传统排序学习不一样的地方：传统排序往往是学习一些人为构造特征（如 TF-IDF 相似度，PageRank 值）的权重，其中的相似度是人为预先定义好的，如 TF-IDF 相似度被定义在余弦相似度上；而对于跨媒体排序，由于异构模态数据之间的语义鸿沟，学习的往往是一个跨媒体映射函数，使得不同模态数据被映射到同一个特征空间，进而再进行语义相关性排序。目前具有代表性的跨媒体排序方法主要有 PAMIR（Passive Aggressive Model for Image Retrieval）和 LSCMR（Latent Semantic Cross-Modal Ranking）等。

“面向公共安全的跨媒体计算理论与方法”973 项目在杭州举行学术研讨会

为了做好 973 项目的中期总结工作，国家重点基础研究发展计划（973 计划）项目“面向公共安全的跨媒体计算理论与方法”（项目编号：2012CB316400）于 2013 年 7 月 4 日在杭州召开了学术研讨会。

本次学术研讨会由项目首席科学家庄越挺教授主持。项目课题负责人赵耀教授、薛建儒教授、黄庆明教授、李学龙研究员、何晓飞教授、于慧敏教授、黄华新教授、蔡莲红教授、兰旭光教授、於志文教授及主要骨干成员等出席了本次启动会。

首席科学家庄越挺教授首先对赵耀教授入选长江学者特聘教授和蒋树强副研究员入选国家自然科学基金优秀青年基金项目等荣誉表示祝贺。庄老师同时要求以课题为单位，对照任务书，认真做好中期总结工作。

赵耀教授首先代表课题一进行了中期总结汇报。课题一围绕跨媒体数据统一表示这一科学问题汇报了两年来在跨媒体基元提取与描述、跨媒体关联性语义结构一致性描述、跨媒体关联建模等三方面的研究工作。跨媒体基元提取与描述是跨媒体数据分析的基础，其目标是用较为精炼的数据形式来描述数量庞大的跨媒体内容，实现跨媒体数据的约减。跨媒体语义结构一致性描述试图将异构数据在各种基元表示空间的描述映射到一个统一的语义共享子空间，使得具有相同语义的异构数据在共享子空间的点重合或邻近。一旦实现跨媒体数据的基元表达及语义一致性描述，就可通过迁移学习方法来构建跨媒体数据间存在的关联关系，以便于利用这些关联关系来更好的理解跨媒体语义内容。

薛建儒教授代表课题二进行中期总结汇报。课题二主要汇报了主被动相结合的跨媒体数据有效性鉴别、跨媒体数据真实性和完整性认证方法、 Who-What-How 属性感知、行为建模与关注度模型构建、群体交互行为与社群发现等方面的进展。

黄庆明教授代表课题三进行中期总结汇报。课题三的汇报主要围绕从传统主题建模到结构化主题建模展开，即通过对跨媒体数据中的“信息和行为”计算，提取结构性主题。前两年课题三主要从跨媒体话题检测、话题追踪、事件相关性度量、基于对象共生网络的跨媒体数据的快速语义标注、基于核化哈希技术的样本选择以及热点话题及事件结构化计算等方面进行了深入研究。在跨媒体语义单元学习、基于多特征的跨媒体语义分析方法、判别模型和字典学习方法以及跨模态相关模型学习等方面取得了进展。

李学龙研究员代表课题四进行了中期总结汇报。课题四的汇报主要围绕跨媒体计算关键算法上所取得突破性进展展开，包括基于平行向量场理论的跨媒体数据处理、海量跨媒体数据的统一建模等方面。在平行向量场学习理论方面，课题四从向量场的角度建立了新的机器学习框架，提出了基于平行向量场的降维方法、基于平行向量场的排序方法、基于平行向量场的多任务学习方法、基于平行向量场的数据流形对齐方法。在主题建模理论方面，课题四研究的主题建模理论是不同类型媒体数据的相同主题建模，这是跨媒体信息挖掘的一个重要问题。

吴飞教授代表课题五进行了中期总结汇报。汇报内容如下：在跨媒体检索方面提出了结合结构先验知识的跨媒体稀疏特征选择、基于多模态字典学习的保持相关性映射、基于隐条件随机场的跨媒体关联性挖掘等方法，在跨媒体排序方面提出了基于单向结构学习的跨媒体排序、基于双向结构学习和隐性模型分解的跨媒体排序等方法，在跨媒体哈希索引方面提出了保持局部距离的样条哈希索引、多模态数据哈希索引、高阶关联关系哈希索引等方法，在跨媒体内容整合方面提出了跨域监督主题建模、主题中的重要性因子提取、跨媒体主题建模等方法。

于慧敏教授代表课题六进行了中期总结汇报。汇报内容包括主题/事件的演变与因果理论模型和搭建跨媒体计算验证平台。课题六的主要进展包括跨媒体话题/事件的迁移学习、跨媒体非完美标注学习、基于概率论辩的因果推理系统和基于主题一致的因果演化网络构建。

出席会议的老师也认真讨论了课题中期总结需要准备的任务和材料。

“面向公共安全的跨媒体计算理论与方法”973 项目杭州研讨会代表合影（2013 年 7 月 4 日）

多媒体领域国际顶尖学术会议 ACM MM2013 录用项目组三篇跨媒体方向长文

多媒体领域国际顶尖学术会议 ACM Multimedia 2013 在会议征文中首次将跨媒体这一术语单列，显示国际同行对这一领域研究逐渐认同。今年 ACM Multimedia 一共收到了 500 多篇长文和短文的投稿。经过严格评审，长文录用率为 20%、短文录用率为 30%。

本项目组三篇与跨媒体相关的长文被录用，涉及跨媒体双向结构学习排序、跨媒体平行向量场表达和噪音标签学习。具体如下：

1、Fei Wu, Xinyan Lu, Yin Zhang, Zhongfei Zhang, Shuicheng Yan, Yueting Zhuang, Cross-Media Semantic Representation via Bi-directional Learning to Rank, *Proceedings of the 2013 ACM*

International Conference on Multimedia (ACM Multimedia, Full Paper, accepted to appear),2013

2、Xiangbo Mao, Binbin Lin, Deng Cai, Xiaofei He, Jian Pei, Parallel Field Alignment for Cross Media Retrieval, *Proceedings of the 2013 ACM International Conference on Multimedia (ACM Multimedia, Full Paper, accepted to appear),2013*

3、Yingming Li, ZhongAng Qi, Zhongfei (Mark) Zhang, and Ming Yang, Learning with Limited and Noisy Tagging, *Proceedings of the 2013 ACM International Conference on Multimedia (ACM Multimedia, Full Paper, accepted to appear),2013*

“Cross-Media Semantic Representation via Bi-directional Learning to Rank”论文提出了基于双向结构学习的跨媒体排序的方法。这一方法将隐空间嵌入（latent space embedding）引入到结构化支持向量机，支持对异构跨模态数据进行相互查询排序。相对于传统的单个（Itemwise）排序和两两（Pairwise）排序方法，结构化支持向量机方法对列表（Listwise）排序结果的损失函数进行优化，从而能够在不同的排序测量函数（Rank Measure function），如 MAP（Mean Average Precision）和 Precision@K 下得到最佳效果。在跨媒体排序中，结构化支持向量机的权重矩阵衡量了两种模态数据特征之间的相关程度。课题组进一步将权重矩阵约束为低秩矩阵，在最大化间隔（Maximum Margin）准则下求解分解后的两个矩阵。求解得到的两个矩阵不仅将两种跨模态的原始数据映射到同一隐空间，同时在此隐空间内对跨模态排序性能进行优化。为了进一步提升排序性能，课题组在结构化支持向量机中引入跨媒体双向排序约束，使得同一个模型同时可以支持两个不同方向的排序，而且能更好地刻画隐空间，提升跨模态检索性能。

“Parallel Field Alignment for Cross Media Retrieval”提出了基于平行向量场对齐的跨媒体检索的方法。这一方法利用流形学习中的平行向量场对不同多媒体数据进行建模表达，并通过流形对齐（manifold alignment）方法来解决跨媒体检索中的语义鸿沟问题。相对于传统的基于联合建模表达（joint model）的跨媒体检索方法，平行向量场对齐方法针对刻画多媒体数据的向量场对齐结果的损失函数进行优化，从而能够在不同的信息检索测量函数（Rank Measure function），如 MAP（Mean Average Precision）和 PR（ Precision-Recall）下得到最佳效果。在跨媒体检索中，流形对齐方法能够建立两种模态多媒体数据特征之间的关联。课题组用梯度场（gradient field）来刻画数据流形上的平行向量场,从而进一步表达多媒体数据，使得流形对齐过程能更好地揭示不同多媒体数据之间的语义关联，提升跨媒体检索性能。

“Learning with Limited and Noisy Tagging”通过结合使用未标注数据空间中包含的信息和多标签空间中包含的信息，提出了一种多标签约束半参数化正则支持向量机方法。这一方法动机来自于如下两个方面的观察：1）多标签空间中包含的信息可以用来有效的对训练集中的噪声标签进行降噪；2）未标注数据空间中的信息添加到训练集中可以更好的反应出整体数据集的分布特点，使得分类训练过程中更准确的学习倒类别之间数据边际分布的几何结构，从而找出更适合的分界面。该方法结合多标签信息和未标注样本信息两个方面设计判别式模型，并且给出了相关的理论推导证明，目前国内外同类算法都只是集中在一个方面。

国际期刊 IJMIR 推出“Cross-Media Analysis”特刊

International Journal of Multimedia Information Retrieval 国际期刊推出“Cross-Media Analysis”特刊。该特刊客座编辑为张仲非教授、庄越挺教授、Ramesh Jain 教授以及 Jia-Yu (Tim) Pan 博士。

这一特刊将涉及到跨媒体高层语义建模、跨媒体摘要生成、跨媒体跨域迁移学习、跨媒体主题建模、跨媒体时序演化与趋势预测、网络空间与物理世界相互映照机理、网络行为认知分析等方面。

具体信息请访问：http://www.fortune.binghamton.edu/CFP_CMA_IJMIR2013.html

在这个特刊通知中，对“跨媒体”进行了如下描述：

Today there are lots of heterogeneous and homogeneous media data from multiple sources, such as news media websites, microblog, mobile phone, social networking websites, and photo/video sharing websites. Integrated together these media data represent different aspects of the real-world and help document the evolution of the world. Consequently, it is impossible to correctly conceive and to appropriately understand the world without exploiting the data available on these different sources of rich multimedia content simultaneously and synergistically.

Cross-media analysis is a research area in the general field of multimedia content analysis which focuses on the exploitation of the data with different modalities from multiple sources simultaneously and synergistically to discover knowledge and understand the world.

Specifically, we emphasize two essential elements in the study of cross-media analysis that help differentiate cross-media analysis from the rest of the research in multimedia content analysis or machine learning.

The first is the simultaneous co-existence of data from two or more different data sources. This element indicates the concept of "cross", e.g., cross-modality, cross- source, and cross cyberspace to reality. Cross-modality means that heterogeneous features are obtained from the data in different modalities; cross-source means that the data may be obtained across multiple sources (domains or collections); cross- space means that the virtual world (i.e., cyberspace) and the real world (i.e., reality) complement each other.

The second is the leverage of different types of data across multiple sources for strengthening the knowledge discovery, for example, discovering the (latent) correlation or synergy between the data with different modalities across multiple sources, transferring the knowledge learned from one domain (e.g., a modality or a space) to generate knowledge in another related domain, and generating a summary with the data from multiple sources.

There two essential elements help promote cross-media analysis as a new, emerging, and important research area in today's multimedia research. With the emphasis on knowledge discovery, cross-media analysis is different from the traditional research areas such as cross-lingual translation. On the other hand, with the general scenarios of the leverage of different types of data across multiple sources for strengthening the knowledge discovery, cross-media analysis addresses a broader series of problems than the traditional research areas such as transfer learning. Overall, cross-media analysis is beneficial for many applications in data mining, causal inference, machine learning, multimedia, and public security.

ACM Multimedia 2013 即将举办“Cross-Media Analysis and Mining”的 Panel

在今年于西班牙巴塞罗那召开的第 21 届 ACM Multimedia 会议将举办“Cross-Media Analysis and Mining”的 Panel 讨论。

项目首席科学家庄越挺教授和课题六组长张仲非教授将参与讨论，同时也邀请了 Alexander Hauptmann(Carnegie Mellon University, USA)、Ramesh Jain (University of California – Irvine, USA)、Alberto del Bimbo (University of Florence, Italy)、Selcuk Candan (Arizona

State University, USA)、Alexis Joly (INRIA, France)等国际知名学者参与讨论。

项目人才培养情况喜人

1. 新增教育部创新团队 1 个：团队负责人赵耀教授，2012 年度
2. 新增教育部长江学者特聘教授 1 人：赵耀教授，2012 年度
3. 新增国家自然科学基金杰出青年基金获得者 3 人：赵季中教授（2013 年度，公示中)、李学龙研究员和何晓飞教授（2012 年度）
4. 新增国家自然科学基金优秀青年基金获得者 2 人：蒋树强副研究员（2013 年度)、於志文教授（2012 年度）
5. 新增教育部“新世纪优秀人才支持计划”入选者 1 人：郭斌副教授（2012 年度）
6. 新增教育部全国百篇优秀博士论文 1 篇：杨易博士，2012 年度，博士论文题目“跨媒体检索与智能处理关键技术研究”，导师为潘云鹤院士和庄越挺教授
7. 新增中国计算机学会优秀博士论文 1 篇：韩亚洪博士，2012 年度，博士论文题目“基于图模型表达和稀疏特征选择的图像语义理解” 导师为庄越挺教授
8. 李学龙研究员获 2013 年度中国科学院青年科学家奖、於志文教授获 2012 年度中国计算机学会青年科学家奖。

项目网站开通以及建立数据共享服务

项目组设计和开发了项目网站（http://www.dcd.zju.edu.cn/cmctm/），网站中包括项目概括介绍、最新动态、规章制度、项目简报、通知通告、项目成果、联系方式和数据共享等栏目。同时，利用项目网站，初步搭建了“跨媒体数据共享服务平台”。在共享平台上，发布了面向食品安全领域的共享数据，包括了从 2012 年 1 月 1 日至 2012 年 12 月 31 日的国内食品安全相关新闻报道、相关微博评论及网络视频等数据，完成了跨媒体食品安全数据集 1.0 与 2.0 版本的整理与构建。其中食品安全数据集 2.0 版本截止到目前为止，总计包含 53260 篇新闻报道，73511 条食品安全相关微博，3673 张图片资源。

国际相关学术热点和动态

1）Google 开放 “知识图谱“重要资源

Google 收购 Freebase 到推出知识图谱（Knowledge Graph）产品，代表 Google 建设实体（Entity）属性数据（也称为 Ontology 或者 Semantic Web）基本成熟，开始进入产品转化期。初期知识图谱产品只是作为类似「百科」的形式进行展示，并未显示出足够的吸引力。这一努力使得 Google 可以在更广泛的领域直接提供精准答案，而不只是作为搜索结果的中转集散地，即用户得到的不再是一些搜索结果的候选链接，而是直接由实体数据提供的精准结果。2013 年 7 月，CMU LTI 与谷歌合作，放出了知识图谱的重要资源：800 万 ClueWeb 文档通过自动标注产生的 110 亿短语。这些短语全部与 Freebase 实体对应，使得目前大家在 ClueWeb 上的字符串和 n-gram 操作，转变成在知识图谱上对实体和概念的操作，对众多应用影响重大。

2）斯坦福大学 David L Donoho 教授获邵逸夫数学科学奖

2013 年度邵逸夫数学科学奖授予大卫．多诺霍，他是美国史丹福大学的 Anne T and Robert M Bass 人文学讲座教授和统计学教授。他对现代数理统计学作出了深远的贡献：他

开创了在有噪声情况的最优统计估计算法;而他又建立了在大数据中实现稀疏表示和复原的高效率技巧。

在过去的半个世纪，计算技术出现了戏剧性的进步，给数理统计学的理论和应用带来了根本性的新挑战。大卫．多诺霍在这个领域举足轻重，他开发了新颖的数学和统计工具，以处理高维大型数据、噪声污染数据等问题。他以严格数学分析为根基，建立了快速、高效且通常是最优的算法。

他的工作中引入了一些重要主题，这些想法已经成为当今许多理论的典范，包括利用稀疏表像描述复杂对像，以及相关的自适应非线性阈值方法；他建立了稀疏性与某些惩罚函数(特别是极小化 L1 范式) 之间的深刻联系。他的许多工作有一个共通的源头，就是如何建立一套算法，以处理有噪声情况下的统计估计。这些算法颇不平凡，克服了从有噪声数据中恢复信息的困难，而又几乎不会损失任何效率或可信性。在这项工作当中，他展现了小波理论的威力，使很多这一类的统计问题得以处理。多诺霍–约翰斯通 (Donoho–Johnstone) 所建立的软阈值算法，已被广泛应用到统计和信号处理之中。

在过去的十五年中，多诺霍以非线性 L1 范最优化方法为基础，发展了一套处理信号和数据的稀疏和多尺度表示理论。这些技术很好地与非结构化方法和冗余字典功能结合，为复杂问题提供降低维数的基本方法。他与着名数学家伊曼纽尔．卡迪斯和陶哲轩一起，为“压缩感知”技术的发展做出了奠基性的贡献。这个方法“一边感知、一边压缩”，利用极少的数据点，却保留恢复正确信号的能力，对复杂信号 (例如图像) 进行压缩和解压缩时，无论从稀疏性还是复原能力的角度，均为更高效甚至最优的算法。这种方法应用广泛，有关研究领域依然非常活跃。